Universität Zürich^UZH

Master's thesis
presented to the Faculty of Arts and Social Sciences
of the University of Zurich for the degree of
Master of Arts

Inferring text comprehension from eye movements in reading using deep neural sequence models

Author: Laura Celina Stahlhut
Student ID Nr.: 15-721-418

Supervisor: Prof. Dr. Lena A. Jäger
Co-Supervisor: David Reich
Department of Computational Linguistics

Deadline: 01.06.2023

Abstract

In recent years, researchers have explored the connection between eye movements and cognitive processes. This thesis investigates the relationship between eye movement patterns while reading and text comprehension. I infer text comprehension using deep neural sequence models. Building upon previous work, my contribution is engineering linguistic features that are based on findings in psycholinguistics and employing a bidirectional Long Short-Term Memory (BiLSTM) model for this task. The results show that linguistic annotation of the stimulus text improves the model performance. When omitting the linguistic features, my model architecture did not outperform previous model architectures. This highlights the effectiveness of linguistic features in tasks that infer cognitive aspects from eye gaze data. Overall, results in this area are not outstanding. Future research should focus on achieving application-relevant results and exploring alternative approaches with respect to input features and model architecture.

Zusammenfassung

In den letzten Jahren gab es Forschung zum Zusammenhang zwischen Augenbewegungen und kognitiven Prozessen. In dieser Masterarbeit wird der Zusammenhang zwischen Textverständnis und den Mustern in den Augenbewegungen beim Lesen untersucht. Ich schliesse auf das Textverständnis mithilfe von tiefen neuronalen Netzwerken. Aufbauend auf früheren Arbeiten besteht mein Beitrag in der Entwicklung linguistischer Features, die auf Erkenntnissen der Psycholinguistik beruhen, und in der Verwendung eines bidirektionalen Long Short-Term Memory (BiLSTM) für diese Aufgabe. Die Ergebnisse zeigen, dass die linguistische Annotation der Stimulustexte die Leistung meines Modells verbessert. Ohne die linguistischen Features übertrifft meine Modellarchitektur frühere Modellarchitekturen nicht. Dies unterstreicht die Effektivität von linguistischen Features bei Aufgaben, die kognitive Aspekte aus Blickdaten ableiten. Insgesamt sind die Ergebnisse in diesem Bereich noch nicht herausragend. Zukünftige Forschung sollte sich darauf konzentrieren, anwendungsrelevante Ergebnisse zu erzielen und alternative Ansätze in Bezug auf Features und Modellarchitektur zu erforschen.

Acknowledgement

Firstly, I would like to thank my supervisor Prof. Dr. Lena A. Jäger for sparking my interest in the field of eye tracking and for supervising this thesis. Secondly, a big thanks goes to David Reich and Patrick Haller for their theoretical advise, helping me overcome technical problems and letting me use their ideas and code to base my own on. Many thanks to David and my sister Fiona for proofreading this text. I want to thank my friends and family for supporting me throughout the time I wrote this thesis as well as throughout my entire studies within the last couple of years and my friends at the institute who wrote their thesis this semester as well, as they proved to be a valuable network of support. Lastly, I would like to thank the Institute of Computational Linguistics at the University of Zurich for providing me with the infrastructure to complete this thesis and for generously providing me with an office space.

Contents

List of Figures

List of Tables

List of Acronyms

BEyeLSTM	Model presented by Reich et al. [2022]
BiLSTM	Bidirectional Long Short-Term Memory
CNN	Convolutional Neural Network
FFD	First fixation duration
FPR	First pass regression
InDiCo	Individual Differences Corpus
LSTM	Long Short-Term Memory
LinguisticEyeLSTM	The model presented in this thesis (section 5.2)
PoS	Part of Speech
RNN	Recurrent Neural Network
SB-SAT	Stony Brook SAT reading fixation dataset
TF-IDF	Term Frequency-Inverse Document Frequency
TFD	Total fixation duration
TPR	True Positive Rate
TTR	Type-token-ratio

1 Introduction

Reading involves deciphering a text and deriving meaning from it. Research shows that the movements of our eyes, which are commonly referred to as a a *windows into the mind and brain*, reflect cognitive processes such as reading comprehension. On the basis of this notion, there have been a few attempts to infer reading comprehension from eye gaze data using various classification approaches including deep neural networks.

Being able to successfully infer reading comprehension from eye gaze data would provide further insight into the cognitive processes underlying language processing and reading comprehension. Furthermore, a reliable model would eliminate the need for the laborious reading comprehension tests that are performed to assess an individual's level of reading comprehension. This way, it would be possible to quickly evaluate the text difficulty of texts on complex topics which are relevant to the general public, such as voting texts. It could also lead to new research about the importance of the documents for estimating reading skill (Augereau et al. [2016]).

While there have been some promising results in related fields, it has been especially challenging to infer text comprehension where the ground truth label is based on the stimulus text at hand, as opposed to general reading comprehension which is assessed using an independent reading comprehension test. Ahn et al. [2020] compiled the publicly available SB-SAT dataset which contains eye gaze data from 95 undergraduate students reading different text passages, along with their achieved scores from corresponding reading comprehension questions. The authors infer the level of a readers text comprehension with a CNN and an RNN network. They use horizontal and vertical fixation location, fixation duration and pupil size as model input. The state of the art in the task of inferring text comprehension on this dataset is held by Reich et al. [2022], who introduced multiple additional features, including linguistic features, and presented BEyeLSTM, a modular neural network architecture consisting of four subnets. Reich et al. [2022] studied four different tasks: General reading comprehension, text comprehension, subjective text difficulty, whether the reader is a native speaker of the text's language. In the text comprehension task, they used the sequence of fixations as an input and predicted whether an individuals text comprehension was above or below the median in the text. They used the same features as Ahn et al. [2020] as input, along with several reading measures and linguistic features. Interestingly, the ablation study in this work showed that omitting the model component that took most of the linguistic features and the reading measures as an input outperformed the full model in many settings. This leads to the question of whether the good performance is due to their network architecture or the input features. Similar tasks have shown that the inclusion of linguistic features is beneficial in tasks that infer cognitive aspects from eye gaze data.

This thesis aims to further explore the relationship between eye gaze patterns while reading and reading comprehension by performing binary classification of text comprehension using the sequence of fixations and linguistic features of the stimulus text as input. Since Reich et al. [2022] presented a promising approach, I'm building upon their approach in many aspects, including the choice of features, the target label, the network structure and the evaluation protocol. My contribution is the incorporation of further linguistic features which are based on research rooted in linguistics and psycholinguistics. In addition to the linguistic features, I'm using the same scanpath features and very similar reading measures to Reich et al. [2022]. Furthermore, I'm extending the preprocessing pipeline such that it works for German stimulus texts. I'm classifying the sequence of fixations using a simple bidirectional Long Short Term Memory (BiLSTM) and I'm evaluating the results in different cross-validation settings (*New Reader*, *New Book*, *New Page*). I'm working with the SB-SAT dataset with English stimulus texts in order to be able to compare my results with the state of the art. Additionally, I'm performing classification on the Individual Differences Corpus with German stimulus texts, which was compiled at the University of Zurich (InDiCo).

The general goal of this work is to reach the current state of the art results of inferring text comprehension from eye gaze data. The research questions that shall be answered are as follows:

- Does linguistic annotation of the stimulus texts improve the results in the task of inferring text comprehension from eye gaze data?
- Does employing a single BiLSTM that takes all features as input yield better results than employing a network architecture consisting of several subnets?
- Is there a difference between the results on the SB-SAT dataset and the InDiCo dataset?

My approach to answer these questions is to preprocess fixation data and stimulus texts of the two datasets, send them through the same annotation pipeline and train the same model architecture with the annotated data of the scanpaths. Additionally,

I perform an ablation study to investigate the influence of the different kinds of features on the result. I compare the results within and across the datasets, as well as to the scores achieved by Ahn et al. [2020] and Reich et al. [2022].
This first chapter introduces the topic and goal of this thesis. Chapter 2 gives information on the theoretical background and related work. Firstly, I will discuss the topics of reading comprehension and aspects which affect the difficulty of a text. This will be helpful in order to construct linguistic features. Then, I will go on and give some background on the topic of eye tracking and I will highlight some ways in which reading comprehension influences the eye movements while reading. Lastly, I will talk about some previous studies that use eye movements as input feature to predict or infer some cognitive aspect, including some studies on the prediction of reading comprehension that my work builds upon. The problem setting of this thesis is described in in chapter 3 and chapter 4 introduces the two datasets used (SB-SAT and InDiCo), including a description of experiment design and some data analysis. Chapter 5 is the core of this thesis; here, I will describe all the steps undertaken to preprocess and annotate the fixation data and stimulus texts of the two datasets in order to prepare the data for classification. I will also describe the model architecture. In chapter 6, I will specify the details of model training and the evaluation procedure. I present the experiment results in chapter 7 and discuss them in chapter 8. This chapter also includes an overview over the specific experiments I performed in the ablation study and a discussion of the results thereof. Finally, I will answer my research questions and talk about the outlook in chapter 9.

2 Related Work and Background

2.1 Text comprehension

2.1.1 The process of text comprehension

Text comprehension is a complex cognitive process that takes place on various linguistic levels. Ehri [1998] split that process into the act of deciphering a text on the one hand and the comprehension of a text on the other hand. They state that a child learns to comprehend before they learn to decipher text since the reading comprehension process is similar to the listening comprehension process which is acquired when a child learns to speak. Proficient readers are able to focus on the meaning of a text while the mechanics of reading, including deciphering, happen in the background without their awareness. The way a reader decodes a text and processes information on multiple levels in parallel can be illustrated on an example: The following three sentences are the beginning of the text *Dickens* from the SB-SAT dataset, where the protagonist talks about their passion for books.

> ”Even then my only friends were made of paper and ink. At school I had learned to read and write long before the other children. Where my school friends saw notches of ink on incomprehensible pages, I saw light, streets, and people.”

Figure 1 depicts the knowledge sources involved in the text comprehension process. At a very basic level, the reader needs to be able to decode the graphemes of the English language in order to understand this text, i.e. we need to be aware that the letter *E* corresponds to the sound [ˈi], the letter *v* corresponds to the sound [v] and so on. In phonetic writing systems such as English or German, this skill requires the following prerequisites: letter familiarity, phonemic awareness and knowledge of how graphemes typically represent phonemes in words (Pikulski and Chard [2005]). Phonemic awareness enables us to hear individual sounds in words Nurjanah [2018]. Secondly, a reader needs to piece together the letters that make up a word and retrieve the possible meanings of the word. Since the Latin alphabet has a limited set of characters and since many languages use the Latin alphabet instead of having a perfectly customized phonetic writing system, the grapheme representation of a sound can be ambiguous between languages and even writing a language. For instance, the word *even (['ivən])* contains the letter *e* twice but it represents two different sounds. For this reason, readers don't only need to be able to recognize individual letters that they can stick together into a word; readers access their vocabulary knowledge from their oral language skills in order to decipher a word correctly. This can be illustrated with an example presented by Pikulski and Chard [2005]. The words *zigzags* and *onychophagia* (nail biting) both appear infrequently in written form. However, *zigzags* appears much more frequently in spoken language than *onychophagia*. For this reason, mature readers have less difficulty decoding the first word. This example shows that readers don't always decode words letter by letter. Ehri [1998] lists five different ways a reader might read words: 1) decoding the letters of a word, 2) pronouncing and blending familiar spelling patterns (a more advanced form of decoding), 3) retrieving sight words from memory, 4) analogizing to words already known by sight and 5) using context cues to predict words. In order to read fluently, we need to understand most of the words in a text instantly (Nurjanah [2018]). We also need morphological awareness. Knowledge of prefixes, suffixes and base words help us to understand and derive word forms. In addition to the phonetic ambiguity described above, a word can have semantic ambiguity. For instance, the adverb *even* is a homonym which is listed with the following meanings in WordNet 3.1 (Fellbaum [1998]): 1) *used as an intensive especially to indicate something unexpected*, 2) *in spite of; notwithstanding*, 3) *to a greater degree or extent; used with comparisons*, 4) *to the full extent.* Additionally, *even* as a noun can be the synonym of *evening*, as a verb it can have the meaning to *become even or more even* and as an adjective, one of the possible meanings is *divisible by two.*

In order to infer a word's meaning correctly, it is vital that the reader has a good understanding of syntax which helps them understand how the words in a sentence relate to one another and how sentences are linked to form a text. Text cohesion means that ideas within and between sentences can be connected. The ability to connect ideas to other ideas in an overall piece of writing is called coherence (Nurjanah [2018]).

Reading and Text comprehension don't stop at understanding the things that are explicitly stated in the text. A big part of meaning of exists as implicit information in a text. Thus, we have to integrate a great deal of word knowledge in order to understand implicit aspects of a text and rhetorical figures such as metaphors. In our example, the reader needs to be able to infer that the concept of *friends made of paper and ink* describes books and not literal friends, otherwise the passage doesn't make any sense. Many other things are hinted at in the text without being literally stated. For instance, the expression *I saw light, streets and people* suggests that the protagonist is able to fully immerse himself in a story to the point where it equally real to him as the real world. We can also suppose with some certainty that the protagonist must be lonely since he doesn't have any real friends, he must be clever since he learned to read before his peers and his tendency to escape into the world of books rather than live in the real world suggests that his own life does not fulfill him in some way. All in all, the three introductory sentences already tell the attentive reader many things about this story, even if they aren't all explicitly stated.

Research suggests that readers build a representation of a text while they are reading in order to keep track of everything they encountered in the story already and that they connect new concepts with what they already know (Rayner et al. [2006]). Many researchers believe that a reader constructs three kinds of text representations: Firstly, A *surface-level* representation which is a verbatim representation of the wording in the text. Secondly, a *textbase* representation, a propositional representation of things explicitly stated in the text. For instance, for the sentence *One Christmas Sempere gave me the best gift I have ever received*, *Sempre* is the agent, *the best gift I have ever received* is the patient and the narrator is the recipient of the action. Lastly, the reader constructs a *mental model*, *discourse model* or *situation model*. Readers track information along multiple dimensions such as protagonist, time, space, casualty and intentionally. The situation model is continuously updated as new sentences are being read. In order to maintain coherence in their mental model, readers have to generate inferences in order to integrate things that are not explicitly stated in the text into the representation (Gernsbacher and Kaschak [2013]).

Evidently, there are many components involved in reading comprehension, including components related to the text, context, and reader (Snow [2002]). There are also some factors which are specific to the participant that influence reading comprehension, for instance their working memory and attention (Nurjanah [2018]). Furthermore, they need to have critical analytic ability as well as the ability to infer (Snow [2002]). There are inter-individual as well as intra-individual differences in the reading skills of children. Factors such as vocabulary and linguistic knowledge, attention, visualization, inferring, reasoning, critical analysis and working memory, motivation, understanding of the goals of reading, discourse knowledge, domain knowledge and cognitive strategy development have an impact on inter-individual differences, i.e. differences between subjects.

Typically, reading comprehension is assessed with question answering tests, recall measures, oral passage reading tests, and cloze techniques (Fuchs et al. [1988]).

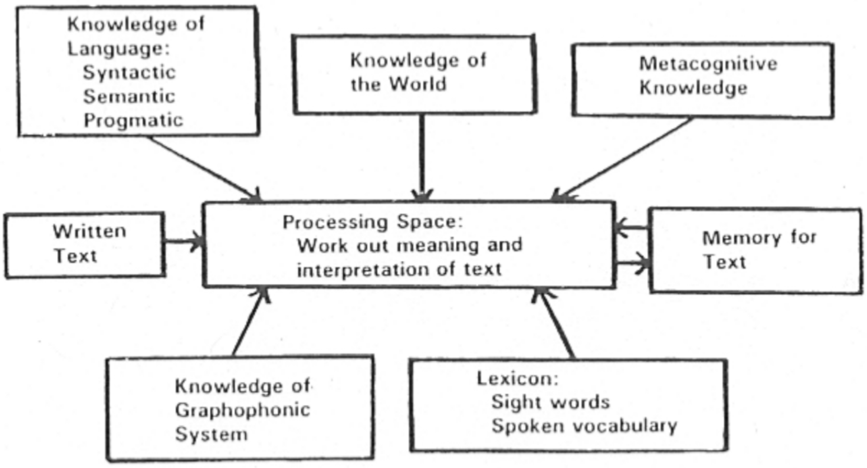

Knowledge of Language: Syntactic Semantic Progmatic
Knowledge of the World
Metacognitive Knowledge
Written Text
Processing Space: Work out meaning and interpretation of text
Memory for Text
Knowledge of Graphophonic System
Lexicon: Sight words Spoken vocabulary

Figure 1: The sources of knowledge involved in text reading and comprehension. Source: Ehri [1998].

2.1.2 Factors influencing text difficulty

As discussed above, various components have an impact on text comprehension, including the features of a given text which can make a text easier or more difficult to understand (Snow [2002]). In order to build a model that can predict reading comprehension, it makes sense to look into some findings as to which aspects of a text have an impact on text difficulty or reading comprehension. This question has been frequently discussed in research, especially with respect to children learning to read and with the desire to locate difficulties in reading comprehension and to improve how we teach children to read (e.g. Nurjanah [2018], Deacon and Francis [2017], Iqbal et al. [2015]).

2.1.2.1 Lexical Properties

On the level of words, some factors regarding the ease with which they can be processed are word length, word structure, word frequency and the part of speech (PoS) of the word.

Longer words take more time to process. Processing time can be an indicator of word difficulty, however, this effect is complex since it interacts with other factors such as word frequency. This so-called word-length effect may be due to some involvement of serial sub-lexical letter processing or due to low-level visual factors that correlate with word length (Barton et al. [2014]).

Words that occur frequently in daily language use are easier to process than low-frequency words. This might be due to the different kinds of reading strategies described in section 2.1.1. High-frequency words can be retrieved by sight from memory while low-frequency or completely unfamiliar words have to be decoded letter by letter. If a word has to be decoded, comprehension might be impacted since the decoding process in fluent reading runs in the background, as discussed above. Marks et al. [1974] showed in their experiment that reading comprehension is significantly higher when 15% of the words in a text are high frequency words as opposed to low frequency words. Naturally, this effect will be bigger if someone has a small vocabulary. Harkio and Pietilä [2016] showed that vocabulary breadth and depth are good predictors of reading comprehension of Finnish learners of English in lower levels of proficiency. Vocabulary breadth and depth in language learners are positively correlated, meaning that learners who had a large vocabulary size also had deeper knowledge of the words (Moghadam et al. [2012]). Since vocabulary size is a key component of reading comprehension (Nurjanah [2018], Rodríguez-Ortiz et al. [2021]), the lexical variation within a text should be considered when determining whether the text is easy or hard to understand. The higher the type-token ratio in a text is, i.e. the more diverse the vocabulary, the more difficult is the text (Klein-Braley [1985]). These effects are interconnected: Stanovich [1986] has found a that a greater vocabulary increases reading comprehension while better reading comprehension also leads to a bigger vocabulary. Similarly, Kieffer and Lesaux [2007] found that there is a reciprocal relationship between vocabulary size and understanding morphology in students.

Another component that enables a reader to read more proficiently is morphological awareness. Morphemes are the smallest linguistic units that carry meaning. They can convey lexical or semantic information of a word, while simultaneously serving grammatical purposes. For instance, the word *incomprehensible* consists of three morphemes: The prefix *in-* conveys negation, the suffix *-ible* denotes the part of speech (in this instance, an adjective) and the morpheme *-comprehen-* is derived from the base form *comprehension.* Morphological awareness enables the reader to derive the meaning of new, morphologically complex words. Research suggests that there is a correlation between morphological awareness and reading comprehension. The analysis of Deacon and Francis [2017] show that morphological structure awareness, morphological decoding and morphological analysis account for 8% of the variance in grade 3 and grade 5 children's reading comprehension.

Another aspect that influences comprehension is part of speech. Jaan [1997] evaluated the relationship between frequency of Parts of Speech (PoS) and reading comprehension and found that a higher percentage of verbs lead to texts being more comprehensible and more interesting. They found that the most frequent PoS in texts were nouns but the impact of the percentage of nouns in a text was unclear. Surprisingly, a high repeating rate of nouns had a negative impact on text comprehension, contradicting the hypothesis above which states that higher vocabulary diversity makes texts easier to understand.

2.1.2.2 Sentence-level Properties

Kadayat and Eika [2020] showed that sentence length is a factor, since in their study, highest comprehension was achieved with sentences comprising 16–20 words. Students with good knowledge on grammar face less difficulty in reading comprehension tests (Nurjanah [2018]) and there is a correlation between a child's level of reading comprehension and their knowledge of syntax in spoken language (Brimo et al. [2018]), however this effect interacts with vocabulary size (Rodríguez-Ortiz et al. [2021]). Due to the significant effect of syntax knowledge on reading comprehension, it can be inferred that texts which are syntactically more complex will be harder to comprehend. Some examples for sentences that might be more difficult to understand are: Sentences with passive verb constructions, adverbial clauses with temporal and causal conjunctions, center-embedded relative clauses and sentences with three or more clauses (Zipoli [2016]). Syntactic complexity can sometimes be governed by text genre. For instance, the argumentative essay requires fairly complex syntactic structures since ideas have to be tied together and linked to opinions, facts and logical reasoning to support arguments (Jagaiah et al. [2020]).

Syntactic ambiguity is another factor which can influence reading behavior and comprehension. While reading a sentence, readers generate hypotheses about the syntactic structure of said sentence which are based on probabilities in general language use. In so-called garden-path sentences, the structure of a sentence is temporarily ambiguous. For instance, in the sentence *The horse raced past the barn fell.*, the word *raced* can be either a past-tense verb or a past-participle. Due to the probabilities in grammar, a reader will most likely assume that *raced* is the main verb, i.e. a past-tense verb. However, once they arrive at *fell*, they will have to correct their initial hypothesis since it becomes apparent that *fell* is the main verb and thus, *raced* is actually a past-participle (Hale [2001]). In such garden-path sentences, readers slow down when the structure is disambiguated in favor of the less preferred alternative (Arehalli et al. [2022]). This finding can be explained by surprisal theory which proposes that the slowdowns are related to the unpredictability of words (Hale [2001], Levy [2008]). Lexicalized surprisal, the words's negative log-probability (Levy [2008]) measures to which extent word's occurrence was unexpected (Frank and Thompson [2012]) based on grammatical probabilities. The higher the lexicalized surprisal, the more unexpected the word. In our example above, the lexicalized surprisal for *fell* would be high. Generally, the difficulty of a word is proportional to its surprisal (Levy [2008]). The amount to which a reader slows down when a sentence garden-paths can be referred to as the cost of a word. Schijndel and Linzen [2021] found that the cost of a word predicted by language models underestimate human garden-path effects. For this reason, Arehalli et al. [2022] propose to calculate syntactic surprisal in addition to lexicalized surprisal. While lexicalized surprisal captures all factors that contribute to a token's predictability (including e.g. word frequency), syntactic suprisal only captures syntactic ones. As such, syntactic surprisal refers to the degree of unexpectedness or surprise associated with a specific syntactic structure within a sentence up until the level of PoS while lexical surprisal, also known as word surprisal, refers to the degree of unexpectedness or surprise associated with a specific word within a sentence.

2.1.2.3 Other Properties

There are various other factors that can impact text difficulty and reading comprehension. On the level of the entire text, they include cohesion and coherence, i.e. the way a text is organized. In a coherent text, ideas are well connected to one another and flow logically with the use of grammatical and lexical cohesive devices Nurjanah [2018].

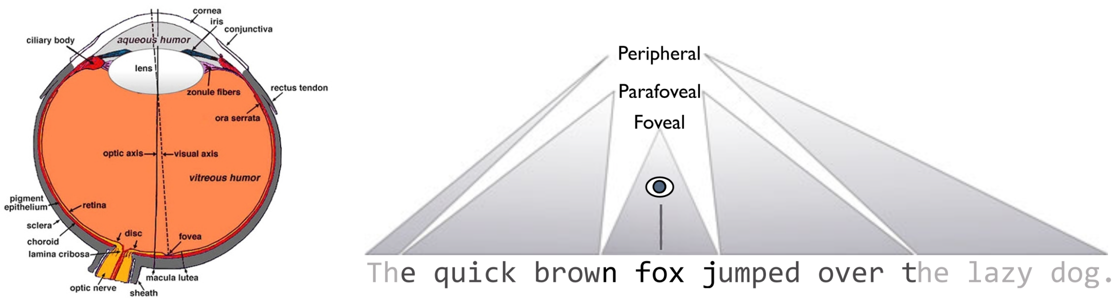
Furthermore, the medium on which the text is presented matters. Electronic text is more challenging to comprehend than printed text, for instance due to the non-linearity if we can scroll. Also, there are distracting factors in digital texts such as hyperlinks, which might distract form reading (Snow [2002]). Another factor is the knowledge and abilities of the reader. Texts vary in topic, genre and discourse style. If the topic is outside of the reader's domain knowledge and/or outside of the reader's interests, a text may be harder to understand to the reader than another one (Snow [2002]).

2.2 Eye movements while reading

Eye tracking makes it possible to record a participant's eye movements while reading. In order to understand how this works, one has to first consider the anatomy and physiology of the human eye which is is illustrated in figure 2. The eye is composed of an external layer (scelera and cornea) which gives the eyeball its white color and refracts incoming light, an intermediate layer (pupil, iris and lens), and internal layer (retina) which is the sensory part of the eye. Light that gets reflected from an object enters the eye through the pupil, which appears black due to the absorption of incoming light by the retina. The iris bundles the amount of light that enters the pupil by contraction and dilation. The lens refracts the light to form a sharp image on the retina, which is is covered by two kinds of photoreceptor cells: Cones, which are color-sensitive, and rods, which are light-sensitive. The area of the retina with the highest concentration of cones is the fovea. This is the central point for image focus, where we have the highest acuity. In the periphoria, where the concentration of cones decreases and the concentration of rods increase, image clarity decreases ([Kolb, 1995]).

Fovea and periphovea don't only denote parts of the internal eye structure; they also name different parts of the visual field which can be categorized into fovea, parafovea and peripheria (see Figure 2). The eye icon in this figure indicates the letter *o* as the location of fixation. The acuity is the highest for the three letters within the fovea, which spans one degree of visual angle around the fixation point. The image is less sharp but letters are still partly recognizable within ten degrees of the fixation point (parafovea). In the rest of the visual field (peripheria), the image is too blurred to recognize letters. Due to the small size of the fovea, the eye has to move in order to perceive visual information, which enables eye tracking technology. The purpose of eye movements is to bring an area of interest into the field of high visual resolution.

Eye movements can be classified into two broad categories: Fixations and saccades. A fixation is a short period of relatively stable gaze and a saccade is a rapid, precise movement between one fixated area and another. Visual information can only be processed during fixations since the quick movement during saccades suppresses vision. Despite the common subjective impression that ones eyes move continuously accross the text when reading, the eyes actually jump from one area to another in a series of fixations and saccades. Words in a text can receive one, multiple or no fixations. It is believed that readers attempt to target the center of words but tend to fall short such that the preferred viewing location is halfway between the beginning and middle of a word (Rayner [2009]). Short and frequent words are more likely to



(a) (b)

Figure 2: (a) Eye anatomy, source: Kolb [1995] and (b) parafoveal processing while reading, source: Schotter et al. [2012]

be skipped than others. Information is being processed during fixations. The eyes of a skilled reader move approximately seven to nine letters per saccade, although such measures vary between individuals (Rayner et al. [2006]).

In left-to-right written languages such as German or English, saccades going to the right or down to the next line are called progressive saccades. Regressive saccades to the left/up, occur frequently during reading whenever a reader goes back in the text, for instance when the reader faces processing difficulties and goes back to an earlier part of the text to make sense of it. About 10%-15% of saccades are regressions (Schotter et al. [2012]).

In video-based eye-tracking-while-reading experiments, infrared light illuminates the pupil and gets reflected on the cornea. This is used to track eye position (Stuart [2022]). Horizontal and vertical locations of fixation are recorded and the fixated word on a screen is identified through the visual angle and distance to screen and camera. Algorithms to classify eye movements into fixations and saccades can be duration- or velocity-based. In research, fixation data is often further processed into reading measures.

The process of visual perception in general, and visual perception while reading, is complex. For instance, it is depicted in Figure 1b that the field of visual perception while reading is shifted towards the right. This means that we can perceive a bigger letter space to the right of the fixation point than to the left. While information to the right of the fixation point is utilized for information processing, information between the currently fixated line is not being utilized in reading. However, in different tasks such as visual search, information below the currently fixated line is being utilized (Rayner [2009]). There are many studies that research foveal and parafoveal processing in order to gain insight in how the human eye processes languages. Back in 1980, the eye-mind assumption had been proposed: ”there is no appreciable lag between what is being fixated and what is being processed” ([Just and Carpenter, 1980, 331]). However, in subsequent research, this theory has often been disproved. Schotter et al. [2012] discuss multiple experiments that illustrate the interplay of foveal and parafoveal processing. An experiment that masked letters in the fovea showed that a great deal of information can be extracted from the parafovea if no foveal information is available. If the letters in the fovea aren't masked, the amount of information retained from parafoveal vision decreases. A lack of parafoveal information disrupts reading to a similar amount as a lack of foveal information. Research suggests that parafoveal information is generally used to decide where to move the eyes and foveal information is used to decide when to move the eyes. The topic of parafoveal-on-foveal effects is being discussed controversially by different researchers (Rayner [2009]). The human eye processes multiple words in parallel and on multiple linguistic levels simultaneously: The decoding of a word takes place at the same time as syntactic parsing (integration of a word into a phrase), semantic processing and referential integration to the broader context.

Many studies show that there is a correlation between eye movement patterns and cognitive processes. As such, eye movements also reflect difficulties readers have while decoding and understanding a text (Rayner et al. [2006]). Reading measures such as the number of fixations and regressions respectively, and total fixation time are an indicator of reading intensity which is related to reading comprehension (Copeland and Gedeon [2013]). Pinnell et al. [1995] showed a correlation between reading fluency and comprehension in fourth-graders. Reading fluency is characterized by accuracy, speed and prosody of oral reading (Pikulski and Chard [2005]). Text difficulty affects fixation duration. For instance, when readers encounter words that are difficult to identify (e.g. low-frequency words or homophones), garden-path sentences or otherwise syntactically complex sentences, fixations get longer (Rayner et al. [2006], Rayner [2009]). Fixation-durations are also longer during mindless reading, when the eye continues to move over across the page despite thinking about something unrelated to the text (Reichle et al. [2010]). Text difficulty also affects saccade length. Saccades become shorter when syntactic ambiguity occurs (Rayner [2009]). Regressions are more frequent in more difficult texts (Rayner et al. [2006]). Short and high-frequency words are more likely to be skipped than long and low-frequency words, however the effect is bigger with word length than word frequency (Rayner [2009]). These effects are more pronounced in poor readers, children that learn to read and dyslexic readers (Rayner et al. [2006].) However, they are not universal, since there are great individual differences when it comes to things like fixation duration, the span of characters the eyes move with each saccade etc. (Rayner [2009]).

2.3 Predicting cognitive processes from eye gaze data

Many studies have leveraged the connection between eye movements and cognitive processes. While psycholinguistic research traditionally treats eye movements as the dependent variable (model output) which is predicted from properties of the reader or text, more recent research has treated eye movements as the independent variable (model input) to infer characteristics of an individual or text (Reich et al. [2022]).
Henderson et al. [2013] and Boisvert and Bruce [2016] have predicted the viewing task from eye gaze data, namely scene search, scene memorization, reading and pseudo-reading and free-viewing, object-search, saliency-viewing and explicit saliency respectively. Kunze et al. [2013] predicted the document type from five different document types. There have been multiple approaches to identify the viewer or reader (Lohr et al. [2020], Makowski et al. [2021], Jäger et al. [2019]) or different traits of the participant such as native language (Berzak et al. [2017]) and personal traits such as extroversion, agreeableness, conscientiousness, neuroticism and openness (Hoppe et al. [2018], Al-Samarraie et al. [2017]). Raatikainen et al. [2021] have used eye gaze data to detect developmental dyslexia in children. Other studies discriminate between cognitive states or cognitive load (Henderson et al. [2013], Shojaeizadeh et al. [2019]).

In these studies, raw eye gaze data is typically converted into sequences of fixations and saccades or further processed into reading measures such as first fixation duration, saccade length and so on. While some of these studies used only scanpath features (Al-Samarraie et al. [2017]) or aggregated eye movement features (Raatikainen et al. [2021]) as an input, the use of aggregated linguistic and gaze data has been a promising approach to predict self-reported language skills, native language, second language proficiency and task classification (Berzak et al. [2017], Berzak et al. [2018], Martínez-Gómez and Aizawa [2014], Hollenstein et al. [2021]).

Methods to predict aspects about a text or viewer form eye tracking data include multivariate pattern analysis (Henderson et al. [2013]), random forests and decision trees (Boisvert and Bruce [2016], Kunze et al. [2013]) and ridge regression (Kunze et al. [2013]). More recently, approaches using deep learning approaches have been successful to identify viewers (Lohr et al. [2020]) and readers (Jäger et al. [2019]). However, it has been challenging to infer higher-level linguistic processes (Reich et al. [2022]).

While there has been a good amount research been done that concerned itself with predicting the behavior of the reader and features of the stimuli, there have only been a few approaches to predicting a person's cognitive state during reading, including text comprehension (Ahn et al. [2020]).

One of the first to attempt predicting reading comprehension from eye gaze data was Underwood et al. [1990]. They used fixations as an input for a multiple regression analysis to classify students into two groups (highly skilled readers and less skilled readers). Ground truth labels were based on a separate reading comprehension test. The authors were able to correctly classify 10 of 15 highly skilled readers and 12 of 15 less skilled readers if they used only mean fixation duration as an input. In their experiments, fixation duration was a good predictor of reading comprehension, but the number of fixations, regressive fixations, reading speed and vocabulary were not. This is interesting since especially vocabulary is often named as an important factor in reading comprehension (Nurjanah [2018], Rodríguez-Ortiz et al. [2021]). The scores on their dataset were good but their model is not generalizeable to an unknown dataset since they trained and tested their classifier on the same dataset (Ahn et al. [2020]).

Augereau et al. [2016] and Lou et al. [2016] used eye gaze data to decode reading behavior and predict an individual's literacy skill with multivariate regression and Support Vector Machines (SVM) with high accuracy. Multivariate regression predicted TOEIC scores with an error of 21.7 points after reading 3 documents and the SVM classification algorithm managed to distinguish high-literacy skilled readers from low literacy skilled readers with 80.3% accuracy. Similarly to Underwood et al. [1990], both papers labeled participants as strong or less strong readers based on a separate reading comprehension assessment, and not based on the stimulus text used in the experiment.

Mézière et al. [2021] use linear regression to predict reading comprehension using different combinations of classical reading measures and reading speed as input. Makowski et al. [2019] develop a model that generates scanpaths in reading from hand-crafted linguistic features of the stimulus text. While they achieved state of the art performance in identifying readers using an SVM with Fisher kernel, they were unable to accurately estimate text comprehension the same way.

Ahn et al. [2020] presented the first attempt to predict reading comprehension from eye gaze data using deep neural networks. They use the raw scanpath, represented by horizontal and vertical fixation location, fixation duration and pupil size as an input to a Convolutional Neural Network (CNN) and a Recurrent Neural Network (RNN). The best result for the prediction of reading comprehension was 65% using the RNN.

As an additional contribution, Ahn et al. [2020] compiled one of the largest datasets of gaze fixations during reading.[1] Therefore, I picked it as one of the two datasets I used to train my models with. A detailed description of the SB-SAT dataset can

[1] The data is publicly available at https://github.com/ahnchive/SB-SAT.

be found in Chapter 4.

Reich et al. [2022] predict i) general reading comprehension, ii) text comprehension, iii) subjective text difficulty and iv) whether a participant is a native speaker of the text they read (i.e. English) from a sequence of fixations and a stimulus text. General reading comprehension is based on the results of a reading proficiency test while text comprehension is based on the answers to comprehension questions belonging to a certain text. Thus, the latter objective is the same as mine. Reich et al. [2022] were the first to approach this problem with a sequence approach that incorporates linguistic features of the stimulus text along with raw fixation data and aggregated reading measures. Linguistic features include word length, lexicalized surprisal, part-of-speech (PoS)-tags, Simplified PoS-Tags, Named Entity Tags, a Content Word feature, the number of syntactic dependents to the left, the number of syntactic dependencies to the right and the distance to the syntactic head. The authors present *BEyeLSTM*, a deep neural network composed of four subnets: *FixNet*, *POSNet*, *ContNet* and *GSFNet*. The first three subnets are composed of two BiLSTM layers and two fully connected layers each while *GSFNet* consists of a dropout layer and a single dense layer. The output of the fours subnets is then concatenated and used as an input to a final dense layer. The different subnets take different types of data as their input. *FixNet* processes the sequence of fixations of the scanpath (x-axis, y, axis, fixation duration and pupil size for each datapoint). *POSNet* processes simlified PoS-features and *ContNet* uses the content word feature as an input. All other linguistic features and the reading measures get aggregated and are fed into *GSFNet*. Reich et al. [2022] use the dataset compiled by Ahn et al. [2020]. Both papers evaluate their model in three cross-validation settings. In the *New page*-setting, individual pages are held out, in the *New book*-setting, entire texts (=5-6 pages) are held out and in the *New reader*-setting, the data from entire readers is held out during training. Results were most promising in the *New Page* evaluation setting, followed by the *New Reader* setting. The *New Book* setting yielded the lowest results, especially in the Text Comprehension and Text Difficulty task. Overall, scores were best in the General Reading comprehension task and in identifying native speakers. Predicting text comprehension was challenging in all evaluation settings. Thus, a desideratum would be to find ways to improve the prediction of text comprehension especially and to find ways to generalize to new books. Reich et al. [2022] currently hold state of the art results in all tasks including the text comprehension task. The code can be found online.[2]

[2] https://github.com/aeye-lab/etra-reading-comprehension

3 Problem Setting

I study the task of inferring text comprehension from a scanpath (i.e. a sequence of fixations) S recorded during the reading of a stimulus text T.[1] Text comprehension is defined as a reader's comprehension level of the stimulus text at the time of recording. The dataset with recordings from N subjects reading M texts is comprised of a set $D = \{(S_{1,1}, T_1, y_{1,1}), ..., (S_{N,M}, T_M, y_{N,M})\}$, where $S_{i,j}$ is the sequence of fixations that has been obtained from the i-th subject reading text T_j; $y_{i,j}$ is the corresponding target label.

The target label is solely based on the scores readers achieved with reading comprehension questions related to the text at hand. This means that the label is not derived from any external reading comprehension assessments. I binarize the number of correctly answered questions into two categories, depending on whether the achieved score is higher or lower than the median score for the current text.

Since the task at hand is framed as a binary classification task, the model's performance can be evaluated with an AUC (area under the ROC-curve). The ROC-curve (receiver operating characteristic) is a probability curve where the True Positive Rate (TPR) is plotted on the y-axis and the False Positive Rate (FPR) on the x-axis. The TPR and FPR are calculated as follows:

$$TPR/Recall/Sensitivity = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

TP, TN, FP and *FN* are True Positives, True Negatives, False Positives and False Negatives. The ROC can be observed if the decision threshold is altered and the AUC is an aggregated measure of performance for all possible classification thresholds. The bigger the area under the ROC-Curve, the better the model's predictions.

[1] Since this thesis is based on the work of Reich et al. [2022], many parts of the problem setting definition have been taken from their paper.

4 Datasets

In this project, I'm working with two datasets: 1) The Stony Brook SAT reading fixation dataset (SB-SAT) with English stimulus texts and 2) the Individual Differences Corpus (InDiCo) with German stimulus texts.

4.1 Stony Brook SAT reading fixation dataset (SB-SAT)

The SB-SAT dataset was created with the aim of developing models that can predict a reader's level of text comprehension by Ahn et al. [2020]. Eye movements of participants were recorded while they read texts and answered comprehension questions. Additionally, participants filled out self-evaluation questionnaires. The data is publicly available on GitHub.[1]

4.1.1 Participants

Ahn et al. [2020] tested 95 undergraduate students (66 female, 29 male). The participants reported diverse native languages, which might have implications for the readers' linguistic processing of the English stimulus texts. 66 identified as native English speakers, 17 as native Chinese speakers, two participants each as native speakers of Korean, Spanish, Mandarin and Punjabi and one participant each reported Italian, Russian, Cantonese, Turkmen and German as their first language.[2] Overall, this corpus tested participants with various first languages, with English native speakers as a clear majority.

[1] https://github.com/ahnchive/SB-SAT

[2] One person noted down both English as well as Chinese as their first language.

4.1.2 Experiment Design

4.1.2.1 Technical Set-up

The authors reported the following specifics about the technical setup of their experiment: The readers' eye positions were recorded with an EyeLink 1000 from SR Research at a sampling rate of 1000 Hz. The reading screens were displayed on a 19-inch flat-screen CRT ViewSonic SVGA monitor with a resolution of 1024 × 768 pixels and a refresh rate of 100 Hz. The monitor subtended a visual angle of 30°× 22°, with the width of three characters spanning approximately 1°.

4.1.2.2 Procedure

After calibration, participants read the four SAT passages described in 4.1.2.3. Each text passage was fully displayed on either five *(Dickens, Northpole)* or six *(Flytrap, Genome)* screens. Participants were able to turn to the next or previous screen or to go to the corresponding comprehension question. After every text, participants answered five comprehension questions and filled out the self-evaluation questionnaire. There was a time limit of five minutes per reading page but there were no time constraints for the screens with the questions. A drift-correction was performed before each trial and re-calibration was done between the texts when necessary.

4.1.2.3 Stimulus texts

In this dataset, there are four stimulus texts: *Dickens*, *Flytrap*, *Genome* and *Northpole*. All texts are practice passages for the *criticical reading* part of the SAT (Scholastic Assesment Test), a standardized test used as entrance exam to colleges and universities.[3] The texts vary in difficulty and genre: *Flytrap* and *Genome* are technical and *Dickens* and *Northpole* are excerpts from works of fiction.[4] More information on the content of the texts can be found in the Table 16 in the appendix. Texts 1 and 4 (*Dickens* and *Northpole*) are displayed on 5 screens each while texts 2 and 3 (*Flytrap* and *Genome*) are displayed on 6 screens each. The general impression I had is that texts 1 and 4 are easier to read than text 2 and 3, both because of the topic and writing style. This is reflected in the text length and mean word length (see figure 3). Text 2 has more words than the other texts, but fewer sentences,

[3] https://collegereadiness. collegeboard.org/sat/practice

[4] *Dickens* is an excerpt of Carlos Ruiz Zafón's *Angels Game* (2008) and *Northpole* is an adaption of *The Balloonist* (1976) by Donald Heiney who wrote under the pseudonym MacDonald Harris.

meaning that it has longer sentences than the other texts. Text 3 has more shorter sentences. The mean word length per text is biggest in the two technical texts, which will probably have implications on the text difficulty as well (see Section 2.1.2).

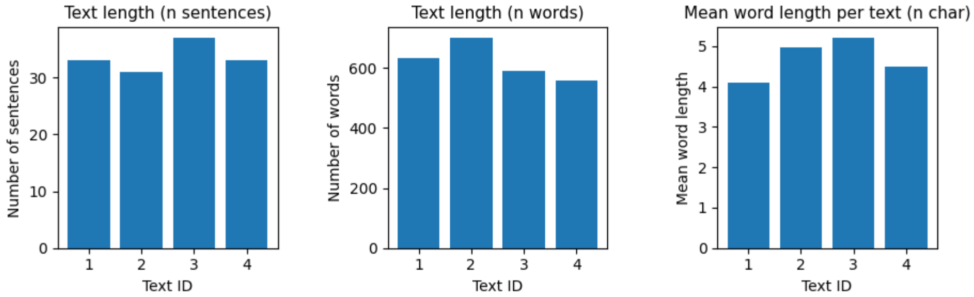
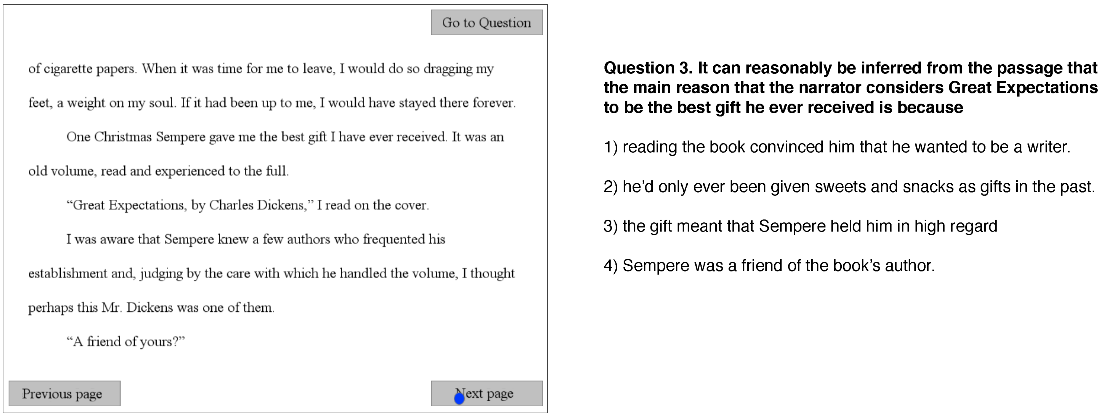


Figure 3: Text length & mean word length of texts. Text IDs correspond to the texts as follows: *Dickens* (1), *Flytrap* (2), *Genome* (3), *Northpole* (4).

Not only text and word length differ from text to text, but also the complexity that comes with the writing style. For instance, in *Dickens*, there is a lot of direct speech and short sentences. In Flytrap, there are many ellipses in the form of brackets or dashes (e.g. in the following sentence: *First, the Flytrap encodes the information (forms the memory) that something (it doesn't know what) has touched one of its hairs*). Another aspect that might lead to lower reading comprehension skills is the familiarity of the reader with the words in the text. Texts 2 and 3 contain many words that readers might not be familiar with (e.g. *transgenic, Biotherapeutics, microcolulombs* etc.). Furthermore, there are some formal features of the texts which might affect reading comprehension. The majority of the screens ends in the middle of a sentence. Sentences being cut off at the end of the screen can disrupt the natural flow of reading (see Figure 4).

Another characteristic of the text might have been confusing for the reader: The unusual feature placement of punctuation marks in direct speech. For instance, in *Genome*, it says: *Welcome to the world of "pharming," in which simple genetic tweaks ...* instead of *Welcome to the world of "pharming", in which simple genetic tweaks ...*. It is possible that this disturbed the reading flow of the participants.

My hypothesis is that such aspects about the text on various linguistic and formal levels have an influence on reading comprehension and can thus been used for my task of predicting reading comprehension from eye gaze data. I will describe my strategy to encode different aspects of the stimulus text as features in Section 5.1.1.2.



of cigarette papers. When it was time for me to leave, I would do so dragging my feet, a weight on my soul. If it had been up to me, I would have stayed there forever.

One Christmas Sempere gave me the best gift I have ever received. It was an old volume, read and experienced to the full.

"Great Expectations, by Charles Dickens," I read on the cover.

I was aware that Sempere knew a few authors who frequented his establishment and, judging by the care with which he handled the volume, I thought perhaps this Mr. Dickens was one of them.

"A friend of yours?"

Question 3. It can reasonably be inferred from the passage that the main reason that the narrator considers Great Expectations to be the best gift he ever received is because

1) reading the book convinced him that he wanted to be a writer.

2) he'd only ever been given sweets and snacks as gifts in the past.

3) the gift meant that Sempere held him in high regard

4) Sempere was a friend of the book's author.

(a) (b)

Figure 4: SB-SAT: Examples of a reading and a question screens from the SB-SAT dataset. (a) *Dickens*, screen 4 and (b) *Dickens*, question 3. Source: github.com/ahnchive/SB-SAT.

4.1.2.4 Comprehension Questions

For each text, there are five multiple choice questions that have four possible answers each, one of which being the correct one. An example of a question screen is depicted in Figure 4. The questions range from general inquiries about the theme of the text to more specific questions regarding textual details. Some of the questions also require the reader to engage in critical thinking and infer meaning beyond what is explicitly stated in the text. It is worth noting that the scores resulting from the comprehension questions don't only reflect how well the reader understood the text, but also how well they understood the question. For instance, question 2 to text *Genome*[5] contains terms that not all readers might understand, especially since not all readers were native speakers of English. Question 5 for *Dickens*[6] shows best that the quality of the questions have an effect on the scores as well: The wrong answers were not chosen well which lead to only 17 participants answering this question correctly, even though the other four questions for this text had some of the highest numbers of correct answers. This suggests that readers understood the text *Dickens* well but didn't understand question 5 for that text well. While keeping these possible

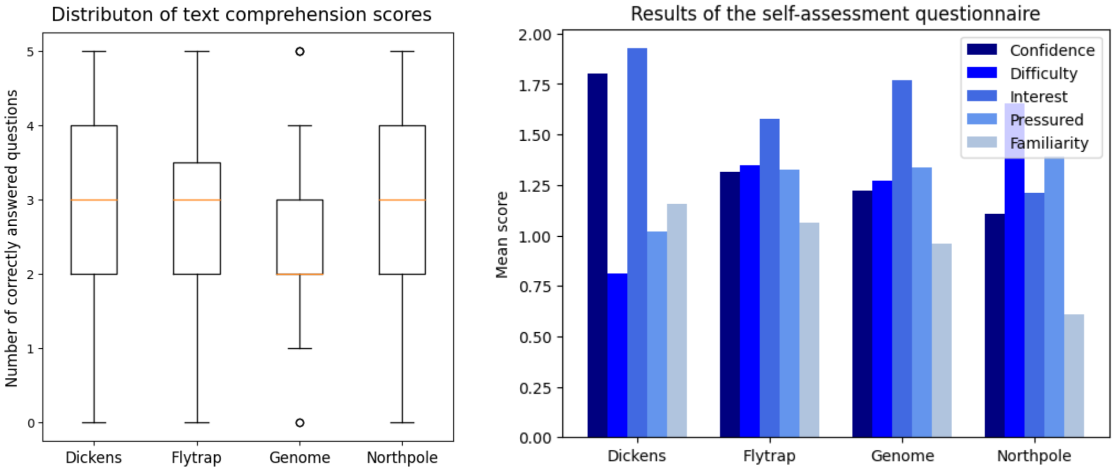
[5] *Genome*, question 2: The authors attitude towards pharming is best described as one of: 1) apprehension, 2) ambivalence, 3) appreciation 4) astonishment.

[6] *Dickens*, question 5: Which statement best characterizes the relationship between Sempre and Charles Dickens? a) Sempre models his own writing after Dicken's style, b) Sempre is an avid admirer or Dicken's work., c) Sempre feels a personal connection to details of Dickens's biography, 4) Sempre considers himself to be Dickens's most appreciative reader. Answer c was counted as the correct answer, but many readers answered b (n=63) or d (n=11), which made sense as well with the given text.

pitfalls in mind, I will nevertheless use the number of correctly answered questions as my target variable.
The participants' answers to the comprehension questions is saved within the fixation dataframe, along with the correct answer to the respective question. As can be seen in Figure 5, the number of correctly answered questions by the different participants range from zero to five in all texts which means that this variable is suitable for a prediction task. The median is two for *Genome* and three for all other texts, which shows that readers had the most difficulty answering questions for *Genome*. The interquartile range is the biggest for *Northpole* and *Dickens*, meaning that the middle 50% of the scores are spread out the most in these texts. Due to the unequal distribution of the number of accurately answered questions, a model may be biased towards predicting the more frequently occurring scores, particularly in the case of the *Genome* dataset where the interquartile range is the smallest.

4.1.2.5 Self-evaluation questionnaire

In the self-evaluation questionnaire which followed each text, participants answered the following questions on their level of confidence while answering questions and on their level of pressure while reading the passage. They had to rate the subjective difficulty of the texts and indicate how interesting it was for them. Furthermore, they answered whether they had read the passage before and how familiar they were with the topic of the text passage. The mean scores of the participants' answers to some of these questions can be found in Figure 5. *Dickens* was perceived as the most interesting and least difficult text. It was recognized the most (n=4). The more technical texts *Flytrap Genome* received similar mean scores, except that people found the latter text more interesting. *Northpole* received the highest score for difficulty despite not being a technical text, which tend to be more complicated. However, the style *Northpole* is posh and old, which probably makes it more difficult. It is interesting that the level of confidence in having given the correct answer varies a lot between the text, even though the median of the number of correctly answered questions is the same in three texts (see Figure 5).
It would be very interesting to incorporate such self-evaluation questions into a model, but since they are only available for this dataset and not for InDiCo, I will refrain from using them in my architecture.



(a)

(b)

Figure 5: (a) SB-SAT Distribution of correctly answered SB-SAT text comprehension questions by text and (b) results of the self-evaluation questionnaire for SB-SAT.

4.1.3 Fixation Data

The raw gaze data was parsed into a fixation report containing the sequence of fixations with the default algorithm by Eyelink which has a velocity threshold of 30°/sec and an acceleration threshold of 8000°/sec^2 [Ahn et al., 2020]. The fixation report[7] is publicly available on Github as a csv-file. The fixation report contains one fixation per line and the interest areas are on the word level. It contains the sequence of fixations for every participant and every text with the following variables:

- Horizontal and vertical coordinates of the fixation: CURRENT_FIX_X, CURRENT_FIX_Y
- Pupil size of the current fixation: CURRENT_FIX_PUPIL
- Duration of the current fixation: CURRENT_FIX_DURATION
- Interest area ID, label, pixel area, run id and dwell time of the current fixation:
  CURRENT_FIX_INTEREST_AREA_ID,
  CURRENT_FIX_INTEREST_AREA_LABEL,
  CURRENT_FIX_INTEREST_AREA_PIXEL_AREA,
  CURRENT_FIX_INTEREST_AREA_RUN_ID,
  CURRENT_FIX_INTEREST_AREA_DWELL_TIME

[7] The file name of the fixation report is *18sat_fixfinal.csv*.

- Previous saccade direction, angle, amplitude, average velocity; whether or not the previous saccade contains a blink and the blink duration:
  PREVIOUS_SAC_DIRECTION,
  PREVIOUS_SAC_ANGLE,
  PREVIOUS_SAC_AMPLITUDE,
  PREVIOUS_SAC_AVG_VELOCITY,
  PREVIOUS_SAC_CONTAINS_BLINK,
  PREVIOUS_SAC_BLINK_DURATION

Additionally, there are variables to identify the subject, book and screen as well as the fixations of the question screens and the answers given for the comprehension questions along with the correct answers. The fixation report contains 463'564 fixations, 263'032 of which belong to reading screens. If all subjects had looked at all screens[8], fixation data should be available for 2090 different combinations of subjects and text screens.[9] However, there are only 2054 combinations of screens and participants because some screens are missing for some participants, e.g. for subject 3, there is only fixation data available for three screens for *Northpole* instead of six. This could either mean that participants skipped screens frequently or that Ahn et al. [2020] deleted some of the sceens from the dataset. The latter option would be unfortunate since I'm interested in using the sequence of fixations for my experiments.

[8] Screens are originally called 'pages' in the SB-SAT dataset

[9] *Dickens* and *Northpole* have 5 screens each, *Flytrap* and *Genome* have 6 screens each. Therefore, fixation data should be available for 95 subjects * 22 screens = 2095.

4.2 Individual Differences Corpus (InDiCo)

The Individual Differences Corpus (InDiCo) was collected by the Digital Linguistics Group at the University of Zurich with the goal to (1) research individual differences in reading, (2) investigate within-subject differences of reading at different daily conditions as well as to (3) assess the cross-methodological measurement reliability in individual differences between eye tracking and self-paced reading experiments. In order to meet this goal, the participants took part in four experimental sessions, two of which were eye tracking (ET) sessions and two of which were self-paced-reading (SPR) sessions. Since I'm only working with the eye tracking data, I will mostly talk about the eye tracking experiments in this section.
I participated in the collection of the data but not in the design of the experiments.

4.2.1 Participants

62 native (Swiss) German speakers (36 female, 26 male) were enrolled in this experiment. However, for 5 of those 62 participants, no eye tracking data is available[10] because they only took part in self-paced reading sessions and 6 participants only completed one of two eye tracking sessions[11]. In my experiments, I was therefore only able to consider the 57 participants who completed at least one eye tracking session. In contrast to SB-SAT, where some of the participants weren't native English speakers, all participants of InDiCo are native German speakers.

4.2.2 Experiment Design

4.2.2.1 Technical set-up

In the eye tracking experiments, eye movements were recorded with an EyeLink Portable Duo from SR Research. A recording PC was operated by the experimenter to control the camera while the experiment was controlled by a presentation PC. Stimuli were displayed in a presentation window with a resolution of 1280 x 1024 px on the presentation display which was positioned 60 cm away from the eye. The distance between the camera and the eye was 45 cm. A keyboard connected to the presentation PC was used by the participants as a response device to answer questions. A chin rest and forehead rest was used in order to keep the head position

[10] Participants 40, 44, 45, 62, 63

[11] Participants 9, 15, 17, 24, 48, 51

constant.

4.2.2.2 Procedure

In the beginning of the first session, participants were briefed about the experiment and asked to sign a consent form which they signed in the beginning of all of the subsequent sessions as well.

In three of the four session, participants completed a number of psychometric tests before reading text passages and answering reading comprehension questions after each passage. In each session, participants read four of the sixteen texts described in section 4.2.2.3. Each text was presented exactly once to every participant in either an eye-tracking or a self-paced reading session. However, since I'm only considering the eye tracking part of the data and not the data acquired in the self-paced reading experiments, in my dataset, every participant will have read only 8 of the 16 texts. The distribution of the texts in my version of the InDiCo dataset can be seen in Figure 6 Each of the text passages was displayed on five screens.

In the eye tracking sessions, participants were first instructed to sit close to the table with a straight back and both feet on the ground such that they were able to hold the position for the duration it took them to read the four text passages. Then, the table height was adjusted. Before the camera set up, participants were asked to answer a number of self assessment questions that were displayed on the presentation monitor (e.g. to assess their level of tiredness using the Karolinska Sleepiness Scale). After the questions, the camera was set up and calibration was done. Then, participants were left to read the text passages. Each text was preceded by self-assessment questions and followed by reading comprehension questions. A drift-correction was performed before each trial which also functioned as a fixation dot such that each trial started at the beginning of the first word on the page. Re-calibration was done between the texts when necessary.

4.2.2.3 Stimulus texts

The stimulus texts for this dataset were taken from mock exams of the TestDaF *Test Deutsch als Fremdsprache* (TestDaF [2000]). TestDaF is a standardized language test for foreign students or researchers wishing to study or work at a German University. It is administered world-wide and is comparable to the international English Language Testing System (IELTS) and the Test of English as a Foreign Language (TOEFL). The exam consists of four parts: Reading, listening, writing and speaking. The reading task has multiple sub-tasks, one of which consists of reading a

text passage and subsequently answering a number of multiple choice questions to assess reading comprehension. The text passages used for this sub-task were written by journalists, university lecturers and scientists. They were originally published in journals, magazines and textbooks (TestDaF [2000]).
The text names, titles, a description of the topic and the genre I assigned to them can be found in the appendix (see Table 15).

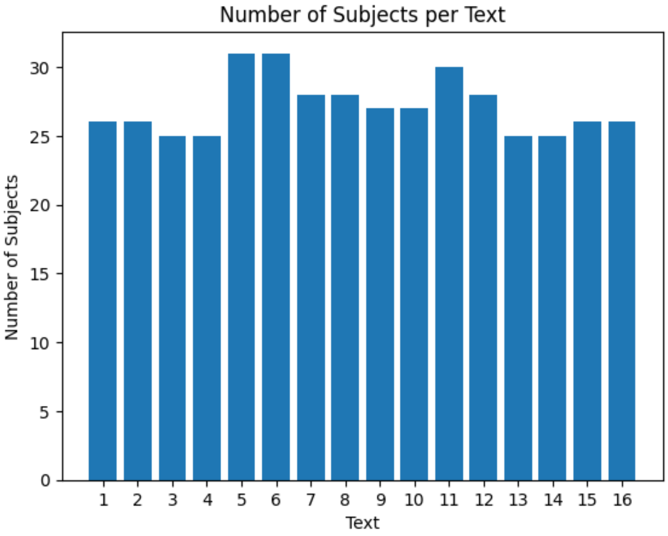


Figure 6: Number of subjects who have read each text in my version of the InDiCo dataset. The number varies because I'm only using a subset of the complete dataset consisting of data from an eye tracking experiment and a self-paced reading experiment. In the orignial dataset, every participand read 16 texts. I'm only using the data from the eye tracking experiments.

4.2.2.4 Comprehension Assessment

After every text, participants answered 10 comprehension questions taken from the original TestDaF mock exam belonging to the text passage. Half of the questions were multiple-choice questions and half of the questions could be answered with *yes*, *no* or *text doesn't say*. Unlike in the SB-SAT experiments, participants were not able to go back to the text once the questions were presented. This strategy provides a better measure for text comprehension at the time of reading. Figure 7 illustrates the distribution of the correctly answered questions across all texts. Most of the participants were able to answer the questions fairly well but there was still some variation between subjects and texts which indicates that this label is likely to be suitable for classification.

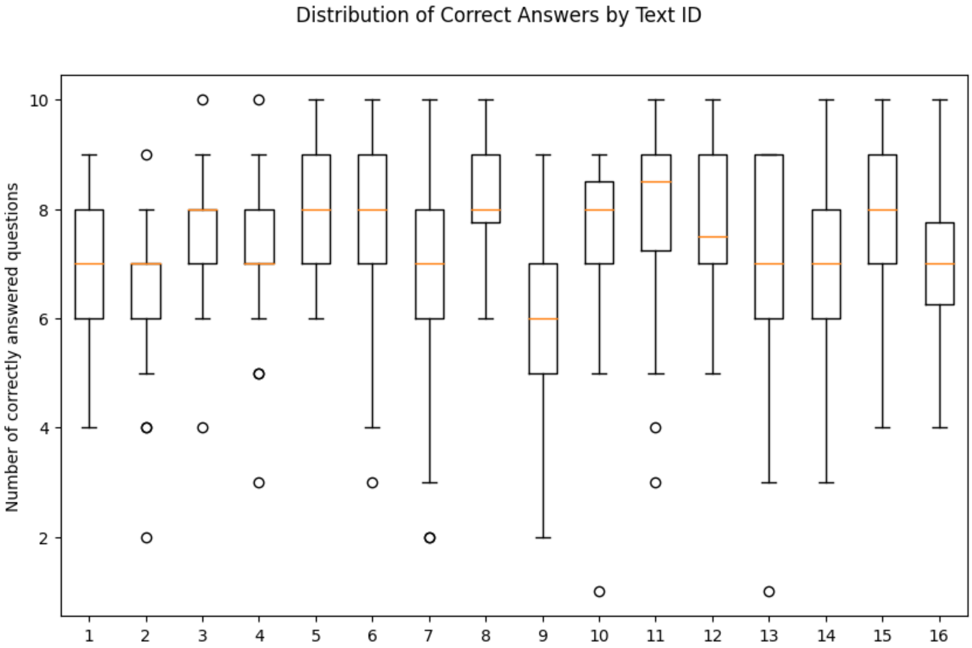


Figure 7: Distribution of the number of correctly answered comprehension questions for the InDiCo dataset.

4.2.2.5 Questionnaire

In every session, participants were asked to answer a number of questions about their condition. The questions surrounded the following topics: Alcohol consumption, eye sight, handedness, hours and sleep as well as medical diagnoses. Participants were also asked to indicate their age and gender. They answered the questions using the keyboard before the trial started.

4.2.2.6 Psychometric Tests

A number of psychometric tests were conducted with every participant in order to assess their cognitive abilities. These tests were used to assess verbal and non-verbal cognitive control (Simon & Stroop, FAIR-2), verbal and non-verbal working memory (sentence span; operation span, memory updating, spatial short-term memory), verbal and non-verbal intelligence quotient (RIAS, MWT-B) as well as lexical an non-lexical reading fluency (SLRT II). It would be very interesting to use the scores from these tests as input features but as the SB-SAT dataset does not have them, I didn't include them for the classification of the SB-SAT dataset, either. Thus, I will not go into further detail regarding the psychometric tests.

4.2.3 Fixation Data

After data collection for the individual differences corpus was concluded, I created a fixation report for all participants with the DataViewer software provided by SR Research. There are many variables available to be exported via DataViewer. In order to ensure that I can process the InDiCo fixations in the same pipeline as the SB-SAT fixations, I included all the same variables in this data that are available in the SB-SAT fixation report (see Section 4.1.3), and a few extra variables in case I needed them later. We collated data from 57 participants who participated in eye-tracking sessions. In every session, they read four texts that were written on 5 screens. However, not all participants participated in both eye tracking sessions. Thus, there are fixations from 2160 screens in this raw fixation report. There is one fixation per line and the interest areas are on the character level. The fixation report gets exported as a tab-separated txt-file.

5 Methods

In this chapter, I will describe my method, including how I preprocessed the stimulus texts and fixation data, engineered features, merged the fixation data with the linguistic features and extracted the target variable. Additionally, I will describe the network architecture. The workflow for the preprocessing steps is depicted in Figure 8. The code can be found in my GitHub repository.[1] I used the existing preprocessing code for the InDiCo dataset as a basis for annotating the stimulus texts, processing fixations and merging fixations with the linguistic annotations.[2] I extended this code to include additional features and to account for the SB-SAT dataset and new participants of the InDiCo dataset.[3] I will declare which parts of my code were taken from existing code and which parts I implemented myself within my scripts as well as within the text of this thesis.

5.1 Data preprocessing and feature engineering

Data preprocessing encompasses the stimulus texts, fixation data and the target label. My goal was to create a pipeline for data preprocessing and model training that works for both the SB-SAT and the InDiCo datasets. However, the raw data of the two datasets was available in different formats, thus the pipeline also includes some dataset-specific prepossessing steps which bring stimulus texts, fixation data and target variable into the same format.

[1] https://github.com/l-stahlhut/inferring-reading-comprehension

[2] https://github.com/hallerp/individual-differences

[3] I obtained the code before the data from all participants of the InDiCo was available. The introduction of new participants required me to account for new types of alignment problems when merging fixation reports with linguistic features (see Chapter 5.1.3).

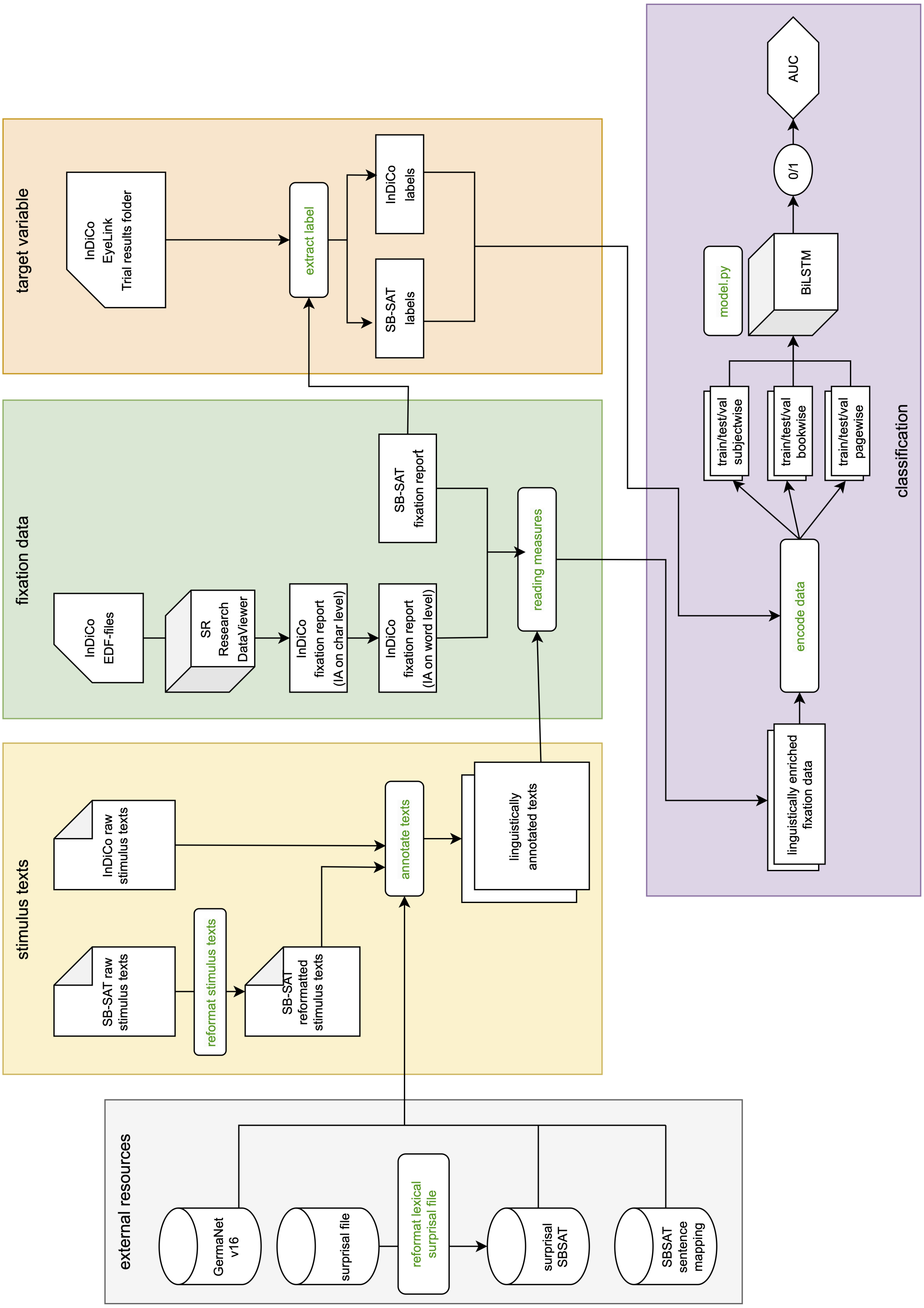


Figure 8: Methods overview: Preprocessing of stimulus texts and fixation data, feature engineering and classification using a BiLSTM. Green represents a python script.

5.1.1 Stimulus texts

The stimulus texts for the InDiCo dataset are available on GitHub as txt-files.[2] There is one file per text screen and one sentence per line. The stimulus texts for the SB-SAT dataset are available as png-files (Ahn et al. [2020]) and as a single tsv-file Reich et al. [2022] that contains one sentence per line along with the information of which text and screen it belongs to.

5.1.1.1 Data cleaning and preprocessing

The first step in preprocessing the stimulus texts is to bring the SB-SAT data into the same format as the InDiCo dataset (one file per screen, one sentence per line). I worked this way around since my preprocessing code is built on the basis of the existing preprocessing code for the InDiCo dataset. To create the individual files per screen from the single tsv-file, I split the sentences by text- and screen-ID and wrote the sentences of all screens to it's separate txt-file.

5.1.1.2 Linguistic Annotation

Once in the same format, the texts of the two datasets were ready to be enriched with linguistic annotations which are based on the findings presented in Chapter 2.1.2. A factor which had to be considered before annotation is that sentences are frequently cut off by the end of a screen in the SB-SAT dataset since they carry over to the next screen. I dealt with this issue during parsing and annotating the stimulus texts: I manually mapped the cut off sentences to the full sentences in a json-file. For the annotation, the full sentences were considered, e.g. for dependency parsing.
The script to annotate stimulus texts is based on a script from the original InDiCo preprocessing code.[2] I adapted the code to include more linguistic features and extended such that it also works for the English stimulus texts of the SB-SAT dataset. The original InDiCo annotation code provides the overall structure of the code and the following functionalities: loading the text screens, parsing the German stimulus texts using SpaCy (Honnibal and Montani [2017]), annotating the texts with dependency tags (number of syntactic dependants on the right and left of the word (**n rights**, **n lefts**), the dependency tag **deps**, dependent tokens on the right and left (**rights**, **lefts**) and **distance to the syntactic head**) and with lexical and syntactic surprisal. It also annotates words with frequency scores but I did not use those scores and opted to calculate them using the wordfreq library (Speer [2022]). The texts are annotated on the word level and results are written to a csv-file.

The code keeps track of the text-ID, screen-ID, sentence-ID and word-ID which is very important since the csv files with the annotated texts later have to be merged with the fixation data which contains fixations on the word level. The original code also contained a class *SurprisalScorer* which I could use to calculate syntactic surprisal on the English texts after adapting the model. All other functionalities were implemented by myself.
I introduced some external data resources in this code, namely GermaNet (Hamp and Feldweg [1997], Henrich and Hinrichs [2010]), a semantic wordnet for German similarly to WordNet, and previously calculated surprisal scores for the SB-SAT dataset obtained from Reich et al. [2022]. As mentioned above, annotations were made on the word level whenever possible since the interest areas of the fixation data which will be used as a model input are also on the word level. German stimulus texts were parsed with SpaCy's transformer model *(de_dep_news_trf)*; the large model *(de_core_news_lg)* was used to obtain semantic similarity vectors and Named Entity features since these features are not available in the transformer model. For English stimulus texts, the transformer model *(en_core_web_trf)* was used to parse texts and the large model *(en_core_web_lg)* was used to obtain semantic similarity vectors. In cases where annotations could only be made on the sentence or text level (e.g. sentence length and text genre) words of the respective sentence or text all received the respective label.

The following features were added by me:
I calculated **word length** in number of characters, **sentence length** in number of words and **text length** in number of sentences. Additionally, I calculated the **mean word length** in each sentence by dividing the number of words in a sentence by the number of characters. I obtained PoS-tags and **Named Entity types** and **-tags** using SpaCy. Following the approach of Reich et al. [2022], I collapsed the PoS-tags into the **simplified POS-tags** *nouns* (N), *adjectives/adverbs* (A) and *function words* (FUNC) and I also implemented their **content word feature**. A mapping from the PoS-tags to the simplified PoS-tags and the content word tags can be found in Table 1.
Since the consensus in the literature states that vocabulary knowledge is a fundamental factor in text comprehension, I added some features which are supposed to capture the lexical variation within a text and the frequency of the words used in a text are in every day language. I calculated **lemma frequency** with the python library *wordfreq*[4] after lemmatization using SpaCy. I also added a feature which indicates whether a word has a **more frequent synonym or homonym** in WordNet or GermaNet depending of the language of the input text. I did this by looking

[4] https://pypi.org/project/wordfreq/

up all synonyms and homonyms of a word in the respective wordnet and calculating word frequencies of all candidate words with the *wordfreq* library and returning True if any of the candidate words were more frequent than the word that occurs in the text. I assume that more complicated texts will have a greater proportion of words that have more frequent synonyms or homonyms. To further measure lexical diversity, I calculated the token type ratio of lemmas in the text (**lemma TTR**) the ratio of content lemmas in the text (**content lemma TTR**) and the ratio of function words in the text (**function lemma TTR**). For every word, I additionally calculated the **TF-IDF** score (Term Frequency-Inverse Document Frequency) to weigh the importance of the term. Term Frequency (TF) measures the frequency fo a word within a document. Inverse document frequency (IDF) measures the significance of a term in the entire collection of documents. The TF-IDF score is the product of the TF and IDF scores, thus it gives more weight to terms that occur frequently within a document but less frequently in the collection of documents.

$$\text{TF} = \frac{\text{number of times the term appears in the document}}{\text{number of terms in the document}}$$

$$\text{IDF} = log\left(\frac{\text{number of documents}}{\text{number of documents containing the term}}\right)$$

$$\text{TF-IDF} = \ \text{TF*IDF}$$

On the level of sentences, I calculated **lexical density**, i.e. the percentage of content words in a text. I also calculated lexical density on the level of the entire text. As a proxy for syntactic ambiguity, I calculated the **pronoun-to-noun ratio** and the **determiner-to-noun ratio** in subsequent sentences. The rational behind this feature is that if there are significantly more nouns than determiners or pronouns in the preceding sentence, the syntactic relations are more likely to be ambiguous. For all consecutive sentences, I also calculated **Semantic similarity between adjacent sentences** using SpaCy and **adjacent sentence overlap**, i.e. the lexical overlap between consecutive sentences to capture aspects such as repetition, expansion, elaboration and coherence. For each sentence, I also annotated whether it was in **active voice** or **passive voice.** However, it should be noted that my function to recognize passive voice depends on the existence of a passivized subject since it is based on SpaCy's dependency tags. It does not capture sentences in passive voice without a passivized subject. For instance *Die Katze wird gejagt* doesn't get recognized by my voice detector but *Die Katze wird vom Hund gejagt* gets recognized.
The tools to calculate *syntactic surprisal* and **lexicalized surprisal** for InDiCo were already present in the preprocessing code. I adapted the model to calculate syntactic surprisal for English stimulus texts and took the surprisal scores from Re-

ich et al. [2022].

In Section 2.1 I mentioned that text format can have an influence on text comprehension. A noticeable difference in the two datasets is that the SB-SAT dataset contains many screens (n=15) with a cut off sentence at the end of the page. This might confuse or irritate the reader. I annotated the words of the affected sentences with the label **sentence is cut off**. The sentences in the InDiCo dataset are never cut off.

Lastly, it was established in Chapter 2.1 that **text genre** can have an impact on text difficulty and reading comprehension. For the annotation of genre, I initially asked ChatGPT to assign a genre to each text without providing categories to speed up the process. I then went back in for a manual correction and decided to use the genres *fiction* and *scientific* for the texts in the SB-SAT dataset and the genres *scientific* and *informative* for the InDiCo dataset. The stimulus texts in the InDiCo dataset all have scientific context (see Section 4.2.2.3). However, not all of the texts originate from academic journals, therefore they vary in style. I made the decision on which genre a certain text belongs to with respect to where the passage was originally published at, the detail in which scientific experiments were described and the general tone of the text.

I annotated the texts with multiple other features, for instance with morphological features from SpaCy. However, I will not further elaborate on annotations that I didn't end up using as features.

I annotated the stimulus texts of both datasets with all features and saved the tokenized and annotated texts as csv-files to later merge with the fixation report on the *text ID*, *screen ID* and *word ID*.

PoS-Tag	Simplified PoS-Tag	Content word	Description of PoS Tag
ADJ	A	True	adjective
ADP	FUNC	False	adposition
ADV	A	True	auxiliary
AUX	FUNC	False	auxiliary
CCONJ	FUNC	False	coordingating conjunction
DET	FUNC	False	determiner
INTJ	FUNC	False	interjection
NOUN	N	True	noun
NUM	FUNC	False	numeral
PART	FUNC	False	particle
PRON	FUNC	False	pronoun
PROPN	N	True	proper noun
PUNCT	FUNC	False	punctuation
SCONJ	FUNC	False	subordinating conjunction
SYM	FUNC	False	symbol
VERB	V	True	verb
X	FUNC	False	other

Table 1: PoS-Tags and simplified PoS-tags. I tagged the texts with spaCy's coarse-grained PoS-tags which follow the universal dependency POS tags. To obtain simplified PoS-tags, I followed Reich et al. [2022] and collapsed the tags into four categories: function words (FUNC), nouns (N), full verbs (VERB) and adjectives/adverbs (A). Words in the content word categories (N, A, VERB) are marked as content words while function words are not.

5.1.2 Fixation Data

5.1.2.1 Data cleaning and preprocessing

Preprocessing raw eye tracking data traditionally consists of denoising data (e.g. removing blinks), event detection (grouping samples in fixations and saccades) and the computation of reading measures based factors such as gaze duration, re-reading time etc. However, as I described in sections 4.1.3 and 4.2.3 I didn't start out with the raw eye tracking data but with the fixation reports generated using the algorithm in the DataViewer software by SR Research. Ahn et al. [2020] exported a fixation report with the reading measures listed in chapter 4.1.3. I exported a fixation report with the same variables from the EDF-files of the Individual Differences Corpus. The tasks of denoising and event detection were thus already taken care of. Nevertheless, it was necessary to perform some data cleaning and re-arranging on the fixation reports before calculating reading measures and merging the fixation data with the lexical features.

SB-SAT fixation data
The SB-SAT fixation report contains the fixations for the reading and question screens from all participants and all texts, one fixation per line with interest areas on the level of words. This data was likely preprocessed already, therefore it required less extensive preprocessing than the fixation report from the InDiCo. Some adjustments were nevertheless necessary: Firstly, I filtered out all lines belonging to a question screen, such that I was only left with screens belonging to reading trials. Then, I added columns for an anonymized subject ID and an ID for the session, which is always 1 in this dataset, since they are also present in the InDiCo dataset. I extracted the subject ID from a column in the data frame containing the recording session. I then deleted lines with fixations outside of interest areas. This includes fixations outside of any lexical material as well as fixations on words which control the flow of the experiments, e.g. buttons to go to the questions or the next or previous page. I corrected the datatypes in certain columns if necessary and I dropped and renamed some columns in order to achieve the same formatting that I ended up with with the prepossessed InDiCo fixation data described below. Finally, I renamed, ordered and dropped some columns and I sorted the dataframe by *subject ID, session ID, text ID* and *screen ID* but not by *word in screen ID* since we're interested in the sequence of fixations and want to preserve regressions.

InDiCo fixation data
The fixation report I obtained from SR Resarch's Dataviewer algorithm is a tsv-file that contains one fixation per line with interest areas on the level of characters.

Data cleaning and mapping the interest areas to the level of words was necessary since the linguistic annotations I will merge the fixation report with are on the word level.

Data cleaning. I read the tsv-file containing the raw fixation report into a dataframe and performed some data cleaning. Firstly, I filtered out rows that only contained NA values and the rows containing the *screen ID* 0 because those are not reading screens. The raw fixation report included several log messages. This lead to multiple cases where the desired information was not in the right cell. An example of such a log message is depicted in table 2. The subject answered *n* (no) to the question of whether or not they had had alcohol on the day of the experiment. In certain rows, the column *READING_TRIAL_ID* contains a log message with the answer to that question instead of the reading trial ID. The first character in the affected cell is the value 3 which actually belongs in the cell indicating the reading trial ID. The last character of the string is the value that is supposed to be in the cell *ALCOHOL_TODAY* and that got replace with *'UNDEFINEDnull'* in these cases. I searched for all affected cells using regular expressions, deleted the log messages and recovered all missing values with the information contained in the log messages.

X	Y	fix. ID	IA LABEL	READING_TRIAL_ID	ALCOHOL_TODAY
731.5	175.6	133	I	1	n
884.5	186.1	144	p	1	n
...	...	...	...	...	...
259.1	508.3	240	e	3 MSG 4103170 !V TRIAL_VAR PARTICIPANT_ALCOHOL_TODAY n	UNDEFINEDnull
356.6	512.8	247	_	3 MSG 4103170 !V TRIAL_VAR PARTICIPANT_ALCOHOL_TODAY n	UNDEFINEDnull

Table 2: Example of log messages and missing values in the exported indico fixation report after donwloading it from SR Research Dataviewer software. X and Y ar the horizontal and vertical coordinates of fixation. This excerpt belongs to the fixation report of participant 12, session1.

Furthermore, I replaced missing values denoted as '.' with NaN values, adapted the data types of the columns when necessary and renamed some columns to match the column names of the preprocessed fixation report of the SB-SAT fixation report.

Mapping of interest areas. I was provided with the code for the transformation of interest areas from the level of characters to the level of words for the InDiCo dataset[5]. This script also adds a new column with the word ID that should correspond to the

[5] https://github.com/hallerp/indiff-preprocessing/tree/main/parsing

word ID in the stimulus text. An example for the mapping of interest areas can be seen in the comparison of tables 11 and 12 in the appendix. I added another column that contains the running count of fixations on a word. I then deleted all rows with fixations that did not fall on an interest area.

Quality Check. As mentioned above, eye tracking data is available for 57 participants who had completed at least one eye tracking session (see section 4.2.1). 51 of the subjects read eight texts while the other six subjects only read four texts because they only participated in one eye tracking session. I manually checked the quality of the fixation data in all 2160 trials of the dataset in the Data Viewer by SR Research. Unfortunately, the overall quality of the collected data is not very good. There are hardly any trials where the recorded fixations fall perfectly on the line. The most common issues include an upwards (sometimes also downwards) drift towards the end of the lines, fixations that are generally too high (multiple lines above the line) and pupils not being recognized by the eye tracker in significant parts of the recording. As a result of my manual inspection, 8 more participants were excluded from the dataset. This means that there are now 43 participants in this dataset, all of which have read 8 of the 16 texts. Unfortunately, the low data quality has implications on the classification results, as discussed in chapter 7.

After these preprocessing steps, I was finally left with one csv-file per dataset containing the sequence of fixations from all trials for the respective dataset with interest areas on the level of tokens and the columns listed in section 4.1.3. These files were then used to calculate reading measures and merge the fixation data with the linguistic annotations described in section 5.1.1.2.

5.1.2.2 Reading measures

Following the approach of many works discussed in chapter 2.3, I transformed the cleaned fixation data into reading measures, which I computed for each interest area in the sequence of fixations, i.e. for each fixated word. The columns included in the SB-SAT fixation report (see section 4.1.3 area a limiting factor on the kind of reading measures that can be derived from the fixation reports. I decided to follow Reich et al. [2022] in the choice of the following reading measures: First Fixation duration (FFD), Total Fixation duration (TFD), Normalized Incoming Regression Count and Normalized Outgoing Progressive and Regressive Saccade Counts. I opted not to implement Words in fixed Context on unigrams, Syntactic Clusters with Universal Dependencies PoS-tags as syntactic labels and Averaged Horizontal and Vertical Location of all fixations due to time constraints on the one hand and the fact that

I didn't use aggregated features in the first place on the other hand. I implemented the calculation of those reading measures myself since I wanted to fit them into my preprocessing pipeline. Additionally, I calculated the number of fixations on a word and first pass regressions. The calculation methods are described below.

First fixation duration (FFD)

First fixation duration is the duration of the first fixation on a word. I calculated this by grouping the fixations of a scanpath by *screen_ID* and *word_ID* and picking the first entry in the column *'CURRENT_FIX_DURATION'*.

Total Fixation Duration (TFD)

Total fixation duration is the sum of all durations of the fixations made on a word within a scanpath. I calculated TFD by grouping the fixations of a scanpath by *screen_ID* and *word_ID* and summing up all values in the column 'CURRENT_FIX_DURATION' on a word.

Number of fixations on a word (n_fix)

I calculated the number of fixations on a word (n_fix) within a scanpath by grouping the fixations of a scanpath and picking the maximum value of the column *'CURRENT_FIX_INTEREST_AREA_RUN_ID'* for each word in the scanpath. *'CURRENT_FIX_INTEREST_AREA_RUN_ID'* is the running count of the number of fixations on a word which was readily available for the SB-SAT dataset and which I calculated specifically for the word-level interest areas for the InDiCo dataset (see section 5.1.2.1).

First Pass Regression (FPR)

For every fixation, I noted down whether it was followed by a first pass regression. To determine FPRs, I first marked for every fixation whether the outgoing saccade was regressive or progressive by comparing the *word ID* of the current fixation with the *word ID* of the next fixation. If the outgoing saccade was regressive and the word was fixated for the first time, i.e. when the *CURRENT_FIX_INTEREST_AREA_RUN_ID* was 1, I noted down a FPR.

Normalized Outgoing Progressive and Regressive Saccade Counts

I summed up the number of outgoing progressive saccades and outgoing regressive saccades from every fixated word and divided that count by the total number of outgoing progressive saccades in the scanpath or outgoing regressive saccades in the scanpath respectively to normalize.

Normalized Incoming Regression Count

Similarly to the detection of outgoing regressive and progressive saccades, I also determined for every fixation whether the incoming saccade was regressive by comparing the *word Id* of the current fixation with the *word ID* of the fixation before. For every fixation, I summed up the total number of fixations on that word and divided by total number of regressive saccades in the scanpath to normalize.

In table 3 and figure 9, I illustrated how I calculated different metrics with the example of the fixations of a participant on the first sentence of the first text in the SB-SAT dataset. The full sentence is "Even then my only friends were made of paper and ink." The participant fixated twice on the words "my" and "friends" but only fixation 3 is a first pass regression since the run ID in fixation 4 is greater than 1. I also indicated outgoing regressive and progressive saccades as well as incoming regressions with which I calculated the reading measures as described above.

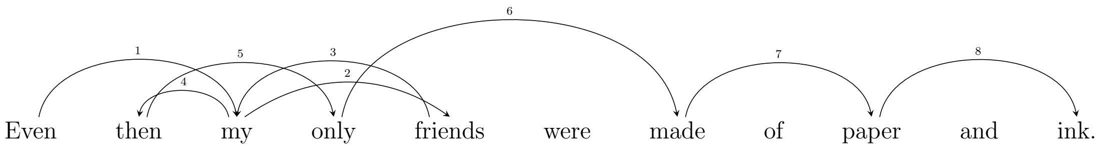


Figure 9: Fixation sequence on a phrase which illustrates regressions. The full phrase is "Even then my only friends were made of paper and ink." and these particular fixations stem from subject 1, text 1, screen 1 (SB-SAT)

fixation ID	word	run ID	fpr	outgoing sac progr	outgoing regr	incoming regr
1	Even	1	0	1	0	0
2	my	1	0	1	0	0
3	friends	1	1	0	1	0
4	my	2	0	0	1	1
5	then	1	0	1	0	1
6	only	1	0	1	0	1
7	made	1	0	1	0	1
8	paper	1	0	1	0	0
9	ink	1	0	0	0	0

Table 3: Illustration of the calculation of first pass regression (fpr). A fpr regression occurs if the run ID is 1 and the word ID is greater than the word ID of the following line. Noted in the field of the fixation after which follows a fpr.

Examples for the reading measures of a scanpath can be found in the appendix for both the InDiCo dataset (table 13) and the SB-SAT dataset (table 14).

5.1.3 Merging fixation data with linguistic features

In order to combine the reading measures of the fixated words with the linguistic features, I merged the two files on the *text ID*, *screen ID* and *word ID* while preserving the sequence of fixations. However, before merging the two files, I had to take care of various kinds of segmentation errors in the files containing the reading measures that lead to alignment problems between the *word ID* of the reading measures and the *word ID* of the lexical features. The root of the problem was that the DataViewer didn't always tokenize the texts correctly when creating the fixation reports. There were three kinds of alignment problems to take care of:

- Alignment problem, type 1: Words that had been split into multiple areas of interest in the fixation report
- Alignment problem, type 2: Multiple words that had been joined together within a single area of interest in the fixation report
- Alignment problem, type 3: Correct mapping between word and area of interest, but wrong value in the *word* column/ encoding errors in the *word* column

To clean the reading measures and merge them with the lexical features, I worked on the basis of a code that was already available for the individual differences corpus.[6]. However, I majorly adapted the code to run more efficiently and to handle cases that appeared in SB-SAT and new subjects in InDiCo.

5.1.3.1 Joining multiple interest areas together

The first type of alignment problem encompassed words that had been wrongly split into multiple areas of interest in the fixation report/reading measures e.g. "Eiszeiten" (W_1) had been split into "Eis" (W_{11}) and "-zeiten" (W_{12}) in the reading measures. As a consequence to this, the *word IDs* of all following words of the same screen (W_{1+1}, W_{1+2}, ..., W_{1+n}, with n = number of words in a screen) were inconsistent with the *word IDs* of the lexical features such that they couldn't be merged. The task at hand was to join the wrongly split interest areas back together. Generally, I merged the interest areas together by replacing the split word with the full word, merging W_2 into W_1, subsequently dropping W_2 and updating all *word IDs* that come after than W_1. However, the concrete operations I made on the different variables differ depending on whether W_1, W_2, both or neither had been fixated, and by the number on fixations on them. Examples of all different scenarios

[6] https://github.com/hallerp/individual-differences/blob/main/src/reading_measures.py

are presented in table 4.

1. W_{11} fixated, W_{12} not fixated

In the case that only the first interest area of the split word was fixated, I adjusted the *word* and *word IDs* as described above and left all other variables as they were. Case 1 is illustrated in table 4: The participant fixated W_{11} ("Mutter-"), but not on W_{12} ("Kind-Situation"). I replaced the split word in fixation 3089 with the merged word (Mutter-Kind-Situation) and adapted the *word IDs* after fixation 3089.

2. W_{11} not fixated, W_{12} fixated

In the case that only W_{12} was fixated, but not W_{11}, I employed the same stratecy as in the first case. An example can be found in table 4: the participant only fixated on "Kind-Situation", but not on ("Mutter-"). Therefore, I replaced the split word in fixation 4278 with the merged word ("Mutter-Kind-Situation") and adapted all the following word IDs to match the word IDs of the lexical features. The reading measures in line 4278 and the following lines were left unchanged.

3. W_{11} fixated, W_{12} fixated

The most complex case to handle was if both areas of interest belonging to the split word had received fixations. In these cases, not only the *words* and *word IDs* were affected by the merge, but also the reading measures. The frequency and order in which W_{11} and W_{12} had been fixated lead to several different ways the problem had to be handled. In these two datasets, participants never fixated more than twice on either of the affected areas. All the different kinds of scenarios that appeared in this dataset are described below with examples in table 4.

Generally, I always checked whether there were any cases of W_{11} and W_{12} being directly fixated after one another. If this was the case, I treated them as two fixations on the same word. In this, my code differs from the original preprocessing code for the Individual differences corpus, since fixations in such cases were merged together, along with the reading measures. However, I would argue that treating them as two fixations on the same word is a better way to handle these cases, since they were recorded as individual fixations on different interest areas belonging to one word. In my case, I did not delete any fixations. I replaced the split word with the joined word in the affected rows and adapted the *word IDs* to match the word ID of the lexical features. I then changed the reading measures as follows: For W_{11}, I set the values for *outgoing progressive* and *regressive saccade on word* to 0. For W_{12}, I replaced the *ffd* with the *ffd* of W_{11} and set the value for *incoming regressive saccade on word* to 0. For both W_{11} and W_{12}, I updated the *n_fix* by summing them up. For cases where there were multiple fixations on W_{11} and/or W_{12} but they were not immediately fixated after one another, I treated them as seperate fixations and

only adapted the *words* and *word IDs* similarly to cases 1 and 2.
Although the principle in how I handled these cases is always the same, I put the different cases present in the dataset in table 4. In *case 3i*, there were two fixations on W_{12} and only one fixation on W_{11}. In *case 3ia*, the reading measures for fixations 1807 and 1808 get adapted because they get treated like two fixations on the same word while the reading measures of fixation 1818 stays the same; only the *word* and the subsequent *word IDs* get adapted. In *case 3ib*, it is the other way around: The first fixation on W_{12} ("-zeiten") gets treated as an independet fixation while fixations 113 and 114 get treated as two fixations on the same word which involves changing the reading measures. In *case 3ii*, on the other hand, there are two fixations on W_{11} and only one fixaiton on W_{12}. The fixations are all directly after one another. They get treated as three subsequent fixations on the same word with a changement in reading measures. In *case 3iii*, there are two pairs of subsequent fixations on W_{11} and W_{12} each. Fixations 5512 and 5517 each get treated as a second fixation on word 82 which changing the reading measures.

4. Neither $\mathbf{W}_{11}$, nor $\mathbf{W}_{12}$ fixated
In case neither of the two problematic interest areas received any fixations, the only measure I had to take was to adapt the *word IDs* after the problematic line. In the example in table **??**, it can be seen that the problematic *word IDs* 82 ("Mutter-") and 83 ("Kind-Situation") were skipped by the reader. I only had to adjust *Word IDs* higher than 82 and was able to leave anything else unchanged.

Case	Subj	Fix. ID	Text	Screen	Word ID RM	Word RM	Word ID corr.	Word corr.
1	28	3089	2	3	82	Mutter-	82	Mutter-Kind-Situation.
1	28	3090	2	3	84	Die	83	Die
1	28	3091	2	3	86	zwischen	85	zwischen
2	41	4277	2	3	81	der	81	der
2	41	4278	2	3	83	Kind-Situation.	83	Mutter-Kind-Situation
2	41	4279	2	3	84	Die	83	Die
2	41	4280	2	3	85	Korrelation	84	Korrelation
3ia	16	1807	2	3	82	Mutter-	82	Mutter-Kind-Situation.
3ia	16	1808	2	3	83	Kind-Situation	82	Mutter-Kind-Situation.
3ia	...	...	...	...	...	...	...	...
3ia	16	1818	2	3	83	Kind-Situation	82	Mutter-Kind-Situation.
3ia	16	1819	2	3	84	Die	83	Die
3ib	4	104	1	2	87	-zeiten	86	Eiszeiten
3ib	4	...	...	...	...	...	...	...
3ib	4	106	1	2	88	werden.	89	werden.
3ib	...	...	...	...	...	...	...	...
3ib	4	113	1	2	86	Eis	86	Eiszeiten
3ib	4	114	1	2	87	-zeiten	86	Eiszeiten
3ii	10	491	2	3	82	Mutter-	82	Mutter-Kind-Situation.
3ii	10	492	2	3	82	Mutter-	82	Mutter-Kind-Situation.
3ii	10	493	2	3	83	Kind-Situation	82	Mutter-Kind-Situation.
3ii	10	494	2	3	85	Korrelation	84	Korrelation
3iii	53	5511	2	3	83	Kind-Situation	82	Mutter-Kind-Situation.
3iii	53	5512	2	3	82	Mutter-	82	Mutter-Kind-Situation.
3iii	...	...	...	...	...	...	...	...
3iii	53	5517	2	3	82	Mutter-	82	Mutter-Kind-Situation.
3iii	53	5518	2	3	83	Kind-Situation	82	Mutter-Kind-Situation.
4	4	152	2	3	78	anders	78	anders
4	4	153	2	3	84	Die	83	die

Table 4: Examples for the correction of segmentation errors of split words in the InDiCo reading measures (Word RM and ID RM are the word and word ID in the reading measures before correction, word ID corr. and word corr are the word and ID after correction to match the ID of of the word in the lexical features.

5.1.3.2 Splitting merged interest areas apart

The second type of alignment problem occurred in cases where multiple words got joined in the same interest area when the fixation report was generated in the DataViewer. The second kind of segmentation error that occurred in the reading measures is that two words were joined together, for example in InDiCo, in one text the two words "hatten.Kleine" were a single interest area, resulting in a mismatch between reading measures word IDs and lexical features word IDs.

In InDiCo, only two tokens were merged together ("hatten.Kleine" and "Pflanzenwelt"), while in SB-SAT, there were quite a few instances of three tokens being merged together.

1. Two words merged together

In some cases, two words were joined together in the reading measures report, but correctly split in the lexical features:

- InDiCo: "zu" + "kanalisieren" (lexical features) vs. "zukanalisieren" (reading measures)
- InDiCo: "hatten" + "Kleine" (lexical features) vs. "hatten.Kleine" (reading measures)
- InDiCo: "Pflanzenwelt" + "Es" (lexical features) vs. "Pflanzenwelt.Es" (reading measures)
- SB-SAT: "them" + "apart" (lexical features) vs. "themapart" (reading measures)

1i. Subject fixated on the word

In case 1i (see table 6), the participant fixated on the merged word. In this case, I updated the word to the first word in the interest area. This is a simplified solution since I don't know which one of the two tokens in the interest area actually got fixated. I changed the word IDs of the following words in the same screen and left all reading measures unchanged. One example from SB-SAT: lexeats [Sempre, &, Sons] vs reading measures [Sempre, & Sons]

- SB-SAT: "species" + "–" (lexical features) vs. "species-" (reading measures)

1ii. Participant didn't fixate on the word In case the participant didn't fixate on the word, the index of the subsequent words have to be moved up. This case is illustrated in table 6: participant 7 didn't fixate the problem area (see fixation ID). Therefore, the only adjustment is the adaption of the word ID after the ID of the problem area.

2. Three tokens merged together

In SB-SAT data, whenever there's a dash in the text, the dash get's merged together with the preceding and following words into one interest area. For instance, the sixth sentence of text 1, screen 1 is: "There was something about them – apart from the letters he could not decipher – that offended him. In the lexical features, the words are tokenized as follows: [..., them, –, apart, ..., decipher, –, that, ...]. However, in the fixation files, the tokens get merged together: [..., themapart, ..., decipherthat,

...].

2i. Subject fixated on the word when multiple lines have been merged into one line, such as when there is a hyphen. It moves the subsequent lines two indexes up and renames the line. If the problem area got fixated, the approach to solve this is the same as in 1i, except that the word IDs of the subsequent words need to be increased by two instead of one (see table 6 where the word id for *from* got increased to 110 from 108.

- SB-SAT: "them" + "–" + "apart" (lexical features) vs. "themapart" (reading measures)
- SB-SAT: "decipher" + "–" + "that" (lexical features) vs. "decipherthat" (reading measures)
- SB-SAT: "change" + "–" + "if" (lexical features) vs. "changeif" (reading measures)
- SB-SAT: "microcolombs" + "–" + "a" (lexical features) vs. "microcolombs" (reading measures)
- SB-SAT: "together" + "–" + "flowed" (lexical features) vs. "togetherflowed" (reading measures)
- SB-SAT: "animals" + "–" + "first" (lexical features) vs. "animalsfirst" (reading measures)
- "voilà" + "human" (lexical features) vs. "voilàhuman" (reading measures)

2ii. Participant didn't fixate on the word If the participant didn't fixate on the problematic word, the only action required was to fix the word IDs of the subsequent words.

Case	Subj	Fix. ID	Text	Screen	Word ID RM	Word RM	Word ID corr.	Word corr.
1i	6	187	4	4	79	zukanalisieren,	79	zu
1i	6	188	4	4	81	regulieren	82	regulieren
1ii	7	643	4	4	52	zunächst	52	zunächst
1ii	7	644	4	4	81	regulieren	82	regulieren
1ii	7	645	4	4	82	und	83	und
1ii	7	646	4	4	83	zunehmend	84	demnächst
2i	1	88	1	1	107	themapart,	107	them
2i	2	94	1	1	107	themapart	107	them
2i	2	95	1	1	108	from	110	from

Table 5: Examples for the correction of segmentation errors of joined words in the reading measures (Word RM and ID RM are the word and word ID in the reading measures before correction, word ID corr. and word corr are the word and ID after correction to match the ID of of the word in the lexical features.

5.1.3.3 Changing name, no ID adjustement

This was necessary in the following problem: text 9, screen 4, wordid 2: lexical features word VHB vs. reading measures word www.vhb.org In these cases, changing the word ID was not necessary, only to rename the words.
In SB-SAT fixation data, there were many encoding errors, especially when there was an apostrophe or quotation marks in the word (e.g. didn't, I'd, company's etc.). For all of these cases, I checked whether alignment with the word ID in the fixation data was correct and then I replaced the word from the fixation data, which had an encoding error, with the word from the lexical features, which didn't have an encoding error. Examples for this are: didnt vs didn't, Id vs. I'd, companys vs company's.

Case	Subj	Fix. ID	Text	Screen	Word ID RM	Word RM	Word ID corr.	Word corr.
1	4	185	9	4	2	www.vhb.org,	2	VHB
1	4	186	9	4	2	www.vhb.org	2	VHB
1	4	187	9	4	3	stehen	3	stehen

Table 6: Examples for the correction of the words in the reading measures dataframe. No other changes than renaming were made here, word IDs were kept as they were

After all these cases were handled, words and word IDs of both files (lexical features and reading measures) were aligned. I joined the two dataframes on the columns subject ID, text ID, screen ID and word ID.

5.1.4 Target variable

The target label is based on the score in the comprehension questions achieved by a specific participant for a specific text. For the SB-SAT dataset, this score can be extracted from the fixation report available on GitHub. For the InDiCo dataset, I took the score from the results folder from the deployed experiment. In the SB-SAT dataset, participants answered five comprehension questions per text, in the InDiCo dataset, they answered ten questions per text. This doesn't matter since firstly, the number of questions within each dataset is consistent secondly, I binarize the scores to get the target variable. I explain in chapter 3 how I transformed the comprehension scores into the target variable.
It is worth noting that the difficulty of the reading comprehension task doesn't only depend on the difficulty of the stimulus text but also on the difficulty of the comprehension questions (see TestDaF [2000]). However, I will work under the assumption that the comprehension questions all have a similar level of difficulty.

5.2 Model Architecture

In figure 10, it is illustrated how the features of a scanpath (described in table 10) are fed into the network. The input to a single input node is a tensor comprised of all features of a single fixation. The number of input nodes corresponds to the number of fixations. The input shape is defined in the first bidirectional LSTM layer (Hochreiter and Schmidhuber [1997]). The model expects a 3D tensor as input (batch size, time steps, input dimension) where batch size is the number of scanpaths in a batch, time steps is the number of fixations in the scanpath and input dimension is the number of features. The input is fed into a single bidirectional LSTM. The network architecture is depicted in figure 10. It is comprised of two bidirectional Long Short-Term Memory (BiLSTM) layers with 75 units each, three dropout layers with a dropout rate of 0.3 and three dense layers with 50, 50 and 20 layers respectively and a tanh activation function. The network concludes with an output layer consisting of a dense layer with a single unit and a sigmoid activation function for binary classification. I arrived at this architecture after tuning the model on the SB-SAT dataset using all scanpath features, reading measures and linguistic features as an input (s1_rm1_lf1) that are listed in table 10. Tuning involved experimentation with the number of layers and input nodes in the different layers, the activation functions, the batch size and the number of epochs the model trained for.

BiLSTM layers are a type of recurrent neural network consisting of two LSTM sublayers that process sequential data in both forward and backward directions. This allows the network to capture bidirectional dependencies in the input sequence, in this case the scanpath. The dropout layers are used to prevent overfitting and the dense layers enable the network to extract features and capture and transform data and produce a final output. I call this model LinguisticEyeLSTM since my focus was to introduce more linguistic features and classify eye gaze data using an LSTM.

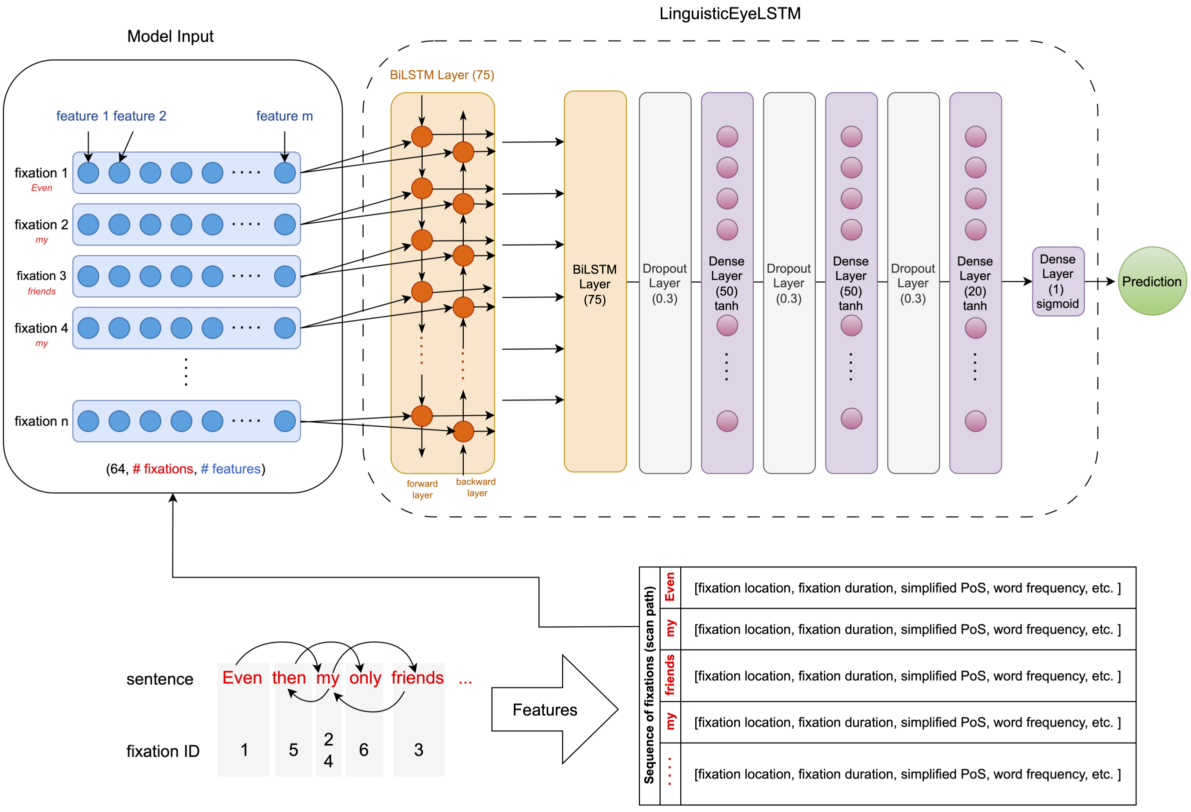

LinguisticEyeLSTM
Model Input
feature 1 feature 2
feature m
fixation 1
Even
fixation 2
my
fixation 3
friends
fixation 4
my
fixation n
(64, # fixations, # features)
BiLSTM Layer (75)
forward layer
backward layer
BiLSTM Layer (75)
Dropout Layer (0.3)
Dense Layer (50) tanh
Dropout Layer (0.3)
Dense Layer (50) tanh
Dropout Layer (0.3)
Dense Layer (20) tanh
Dense Layer (1) sigmoid
Prediction
sentence
Even then my only friends ...
fixation ID
1 5 2 4 6 3
Features
Sequence of fixations (scan path)
Even
[fixation location, fixation duration, simplified PoS, word frequency, etc.]
my
[fixation location, fixation duration, simplified PoS, word frequency, etc.]
friends
[fixation location, fixation duration, simplified PoS, word frequency, etc.]
my
[fixation location, fixation duration, simplified PoS, word frequency, etc.]
[fixation location, fixation duration, simplified PoS, word frequency, etc.]

Figure 10: Model architecture. Scanpath features, reading measures and linguistic features are assigned to the sequence of fixations and used as the model input. In total, there are $M = 37$ different features. An overview of the linguistic features can be found in table 10 One scanpath is comprised of the fixations of one participant within one page of a stimulus text. The model infers whether reading comprehension for a specific scanpath were above or below the median for the respective stimulus text.

6 Experiments

6.1 Model training

I trained LinguisticEyeLSTM with the input features listed in table 10. All models were trained with GPU support using the Keras and Tensorflow libraries. I used an Adam optimizer with a learning rate of 0.001 and set training for 1000 epochs with a patience of 50. Effectively, the model ended up training for ca. 50-80 epochs depending on the dataset and amount of input features used. The code is available on GitHub.[1]

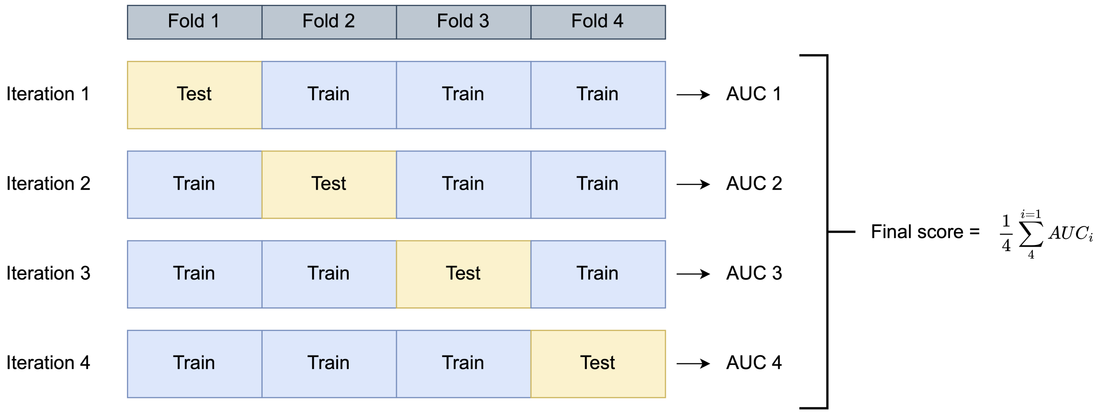
6.2 Evaluation procedure

To evaluate the performance of my models, I used the area under the ROC curve and the three cross-validation settings that have also been explored by Reich et al. [2022] and Ahn et al. [2020]. Cross-validation is a method employed to assess the performance and generalization ability of a predictive model. The dataset is split into multiple subsets (”folds”) and the model is iteratively trained with a different fold being used as the training set each time. To obtain a final score, the scores of the different folds are averaged. The standard deviation of the different scores indicates how well the model is generalizeable across different datasets. I employed three cross-validation settings that vary in how the dataset is divided to form the training and test set in the different folds:

In the **New Reader setting**, fixation data from certain readers is held out during training. For readers that are not held out during training, the model sees all books and all pages during training. In the **New Book setting**, entire books are held out during training. The model sees all pages and fixations of all participants who read the book in the books that are used during training. In the **New Page setting**, individual pages are randomly held out during training. Of the pages that are not held out during training, the model sees fixations from all participants.

[1] https://github.com/l-stahlhut/inferring-reading-comprehension

Four-fold cross validation is performed on the New Book setting and five-fold cross validation is performed on the New Reader and the New Page settings. In each evaluation setting, a train or test instance consists of the sequence of fixations (scanpath) recorded from a participant reading one page of text. I follow the same evaluation protocol for InDiCo. The code for my experiments is based on the code by Reich et al. [2022].[2] I changed the model architecture and adapted the training and evaluation procedure to work for the new model and the problem setting described in chapter 3.



	Fold 1	Fold 2	Fold 3	Fold 4	
Iteration 1	Test	Train	Train	Train	⟶ AUC 1
Iteration 2	Train	Test	Train	Train	⟶ AUC 2
Iteration 3	Train	Train	Test	Train	⟶ AUC 3
Iteration 4	Train	Train	Train	Test	⟶ AUC 4

Figure 11: Illustration of 4-fold cross validation: The dataset is split into four folds and the model is iteratively trained with a different fold being used as the test set each time. The final result is the average of the results of the different folds, in this case the average area under the ROC cuve (AUC). Not depicted in the figure is the validation set, which is 10% of the training data.

[2] https://github.com/aeye-lab/etra-reading-comprehension.

7 Results and Discussion

7.1 Initial experiments with all features

An overview of the results of LinguisticEyeLSTM taking all features as input for both datasets can be found in table 7. In order to determine the performance of my overall approach, the AUC scores these experiments can be compared to Reich et al. [2022]'s BEyeLSTM and the different approaches presented by Ahn et al. [2020], which are presented in the same table. BEyeLSTM with all four subnets previously held the state of the art in the *New Reader* setting while BEyeLSTM without GSFNet held the state of the art in the other two evaluation settings. This suggests that simplified PoS-tags and the content word feature of a fixation sequence were good predictors of reading comprehension in the *New Page* and *New Book* settings while the inclusion further aggregated linguistic features and reading measures lead to better results in the *New Page* setting. The results of Ahn et al. [2020] were outperformed by Reich et al. [2022] in all evaluation settings.

On the SB-SAT dataset, LinguisticEyeLSTM with all scanpath features, reading measures and linguistic features outperformed BEyeLSTM in all evaluation settings, although only by a very small margin in the *New Page* and *New Book* settings and with a bigger standard error in the *New Page* and *New Book* settings, suggesting that the approach isn't genrealizeable to new data as well in those settings. The results suggest that the inclusion of further linguistic features might have had a positive impact on the prediction of text comprehension from eye gaze data. However, the effect of the input features interacts with the effect of the model architecture. In order to differentiate between the effects of the two factors, the results of the other experiments have to be investigated.

The the influence of the two different datasets can be examined by comparing the results on the InDiCo dataset with my results on the SB-SAT dataset. The results for the InDiCo dataset in both settings were worse than the results for the SB-SAT dataset in all evaluation settings and they didn't exceed chance level (0.5) by much in any setting. I argue that this is most likely largely due to the issue regarding the data quality explained in section 5.1.2.1. The effect of the data quality is likely to overshadow any other effects that might be due to the factors which differentiate the two datasets such as the language or the distribution of texts across subjects. However, other factors might influence the difference in the results, such as the choice of the label as discussed in section 5.1.4 which depends on the quality and standardization of the comprehension questions.

Model	Features	Dataset	New Page	New Book	New Reader
LinguisticEyeLSTM	all	SB-SAT	**0.607 ± 0.031***	**0.524 ± 0.024**	**0.581 ± 0.02***
LinguisticEyeLSTM	all	InDiCo	0.52 ± 0.018	0.506 ± 0.012	0.515 ± 0.02
BEyeLSTM (Reich et al. [2022])	all	SB-SAT	0.596 ± 0.012*	0.504 ± 0.015	0.542 ± 0.015*
BEyeLSTM w/o GSFNet (Reich et al. [2022])	reduced	SB-SAT	0.597 ± 0.013*	0.522 ± 0.013	0.521 ± 0.029
RNN (Ahn et al. [2020])	scanpath	SB-SAT	0.571 ± 0.01	0.507 ± 0.01	0.514 ± 0.024
CNN (Ahn et al. [2020])	scanpath	SB-SAT	0.538 ± 0.006	0.493 ± 0.009	0.485 ± 0.016
Regression (Ahn et al. [2020])	scanpath	SB-SAT	0.539 ± 0.007	0.492 ± 0.013	0.532 ± 0.016

Table 7: AUC results ± standard error of the experiments with LinguisticEyeLSTM with all features in comparison to the results from prior studies. The asterisk * indicates values that are significantly higher than random guessing. The colors indicate whether the score is higher (green) or lower (red) than the state-of-the-art held by Reich et al. [2022] for the model using all linguistic features and reduced linguistic features respectively. For every evaluation setting, previous state-of-the-art is underlined, while current state-of-the-art is shown in bold.

7.2 Ablation study

In this section I will investigate how the different parts of my method contribute to the overall performance of the proposed machine learning model.

7.2.1 Input ablation study

I trained six models per dataset, including the models discussed above that took all features as input. I used the same model architecture for all experiments (see image 5.2) but varied the input features. Following Ahn et al. [2020], I used vertical and horizontal coordinates, pupil size and current fixation duration as features to represent the scanpath. The reading measures and some of the linguistic features are inspired by Reich et al. [2022]. Table 10 provides an overview of all features

used as model input in the different experiments.
Firstly, I trained a model where the input is the same as in BEyeLSTM without GSFNet presented by Reich et al. in order to evaluate the effect of my model architecture:

- **Reduced linguistic features**: Binary classification using LinguisticEyeLSTM with the scanpath features, simplified PoS and the content word feature.

Secondly, I trained the following models to investigate the impact of different types of features on the model performance:

- **No linguistic features**: Binary classification using LinguisticEyeLSTM with all features described in table 10 except for the linguistic features.
- **Scanpath features**: Binary classification using LinguisticEyeLSTM with only the four scanpath features as model input.
- **Reading measures**: Binary classification using LinguisticEyeLSTM with only the reading measures as model input.
- **Linguistic features**: Binary classification using LinguisticEyeLSTM with only the linguistic features as model input.

7.2.2 Results ablation study

The two factors under investigation are the model architecture and the input features.
The ideal way to determine the role of the model architecture would have been to train BEyeLSTM with my input features. However, this was not possible due to time constraints. Therefore, I opted to approximate this approach by training my model with the same input features used in BEyeLSTM w/o GFSNet (Reich et al. [2022]). The input might vary slightly due to different preprocessing methods and independent construction of the features. While I used all features as input to a single BiLSTM, the input features are distributed across three different subnets in BEyeLSTM w/o GFSNet: (1) horizontal and vertical fixation coordinates, pupil size and fixation duration, (2) simplified PoS-tags and (3) the content word indicator. The results of both network architectures with full and reduced input features can be found in table 8. The results on the SB-SAT dataset show that LinguisticEyeLSTM with reduced linguistic features only surpassed the scores of BEyeLSTM w/o GSFNet in the *New Reader* setting, indicating that Reich et al.'s modular architecture consisting of multiple subnets generally outperforms my simple BiLSTM.

While BEyeLSTM outperforms BEyeLSTM w/o GFSNet only in the *New Reader* setting, LinguisticEyeLSTM with all features outperformed the version with reduced features in all settings. This reinforces the notion that the added linguistic features were helpful at improving the results for the task at hand.

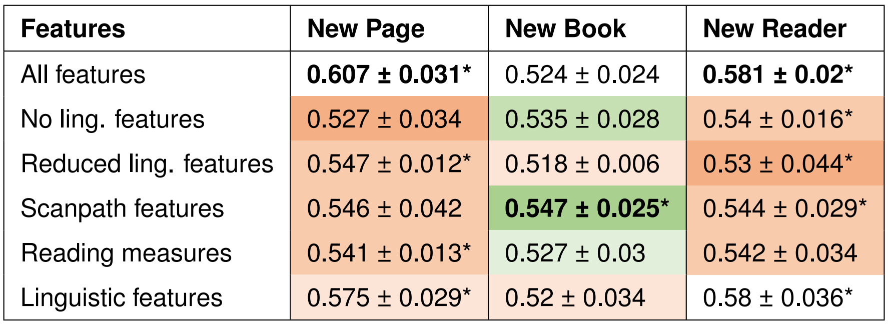
Model	Features	Dataset	New Page	New Book	New Reader
LinguisticEyeLSTM	all	SB-SAT	**0.607 ± 0.031***	**0.524 ± 0.024**	**0.581 ± 0.02***
LinguisticEyeLSTM	reduced	SB-SAT	0.547 ± 0.012*	0.518 ± 0.006	0.53 ± 0.044*
LinguisticEyeLSTM	all	InDiCo	0.52 ± 0.018	0.506 ± 0.012	0.515 ± 0.02
LinguisticEyeLSTM	reduced	InDiCo	0.508 ± 0.019	0.509 ± 0.014	0.515 ± 0.007*
BEyeLSTM (Reich et al. [2022])	all	SB-SAT	0.596 ± 0.012*	0.504 ± 0.015	0.542 ± 0.015*
BEyeLSTM w/o GSFNet (Reich et al. [2022])	reduced	SB-SAT	0.597 ± 0.013*	0.522 ± 0.013	0.521 ± 0.029

Table 8: AUC results ± standard error of the experiments that included all or reduced linguistic features in the input with the state-of-the-art results for comparison. The asterisk * indicates values that are significantly higher than random guessing. The colors indicate whether the score is higher (green) or lower (red) than the current state-of-the-art for the model using all linguistic features and reduced linguistic features respectively. For every evaluation setting, previous state-of-the-art is underlined, while current state-of-the-art is shown in bold. In the case of LinguisticEyeLSTM *all features* refers to scanpath features, linguistic features and reading measures while *reduced* refers to scanpath features, simplified PoS-tags and the content word feature, i.e. the same features used as in BEyeLSTM w/o GSFNet.

Tables 9 and 10 contain the results for the ablation study for the two datasets. Table 10 gives information on the exact features that were used as an input for any given experiment.

The results on the SB-SAT dataset in table 9 show that the best results were achieved if all available input features were used in both the *New Page* and the *New Reader* settings while for the *New Book* setting, the best result was achieved by the model that was only trained with the scanpath features.

In the *New Page* setting, omitting the linguistic features lead to the worst performance. Reducing linguistic features, only using scanpath features or only using reading measures all lead to similarly mediocre results. Using only linguistic features was slightly better but this version of the model could still not reach the performance of the model that used all features. Using a combination of scanpath and linguistic features as well as reading measures proved to be beneficial in this setting, with a strong emphasis on the linguistic features.



Features	New Page	New Book	New Reader
All features	**0.607 ± 0.031***	0.524 ± 0.024	**0.581 ± 0.02***
No ling. features	0.527 ± 0.034	0.535 ± 0.028	0.54 ± 0.016*
Reduced ling. features	0.547 ± 0.012*	0.518 ± 0.006	0.53 ± 0.044*
Scanpath features	0.546 ± 0.042	**0.547 ± 0.025***	0.544 ± 0.029*
Reading measures	0.541 ± 0.013*	0.527 ± 0.03	0.542 ± 0.034
Linguistic features	0.575 ± 0.029*	0.52 ± 0.034	0.58 ± 0.036*

Table 9: AUC results ± standard error for training LinguisticEyeLSTM on the SB-SAT dataset with different features. The asterisk * indicates values that are significantly higher than random guessing. The colors indicate whether the score is higher (green) or lower (red) than the approach that used all linguistic features. Bold numbers highlight the best results in a given evaluation setting.

In the *New Reader* setting, the results look similar in many cases: Omitting or reducing linguistic features, using only scanpath features or using only reading measures lead to mediocre results. Interestingly, the results of the model that was trained with linguistic features only was on par with the model trained with all features in the *New Reader* setting. This is the setting where I was able to achieve better results than the state of the art quite consistently. The model seems to be able to generalize well to new readers when the stimulus text is linguistically annotated, which implies that a reader's reaction to different properties of the text varies strongly.

In the *New Book* setting, the pattern looks different from the other two settings. Here, the best results were achieved when only using the scanpath features, followed by the combination of scanpath features and reading measures and only reading measures. In fact, the inclusion of linguistic features consistently leads to worse results in the *New Book* setting. The linguistic features and reading measures were not helpful when generalizing to books that were not seen during training. That being said, the results in the *New Book* setting can't compete with the results in the other setting in general. In fact, the results of the model that only used scanpath features are comparable across settings, while the results with any other input is just much worse in the *New Book* setting than in the other two settings.

The ROC curves in figure 12 also illustrate how the model's performance in the *New Book* setting is excelled by the performance in the other two settings, a pattern which is also found in Reich et al. [2022]'s ROC curves. The only improvement that can be seen when comparing their ROC curves with mine is in the *New Reader* setting. The plots also show that the model benefits from linguistic features, especially in the *New Page* setting but also in the *New Subject* setting. In the latter, using only reading measures pretty much matches the results of using all features. In the *New Reader* setting, chance level is barely outperformed by any combination of input features.

Comparing figure 12 to figure 13, it becomes apparent that the classification results on the InDiCo dataset don't exceed chance level in any setting. I conducted the ablation study with this dataset despite the disappointing results in the initial experiments which are most likely due to the data quality issues as mentioned above. Despite the low AUC-scores, the patterns in the results displayed in table 10 are interesting since they are similar to the patterns in the results for the SB-SAT dataset

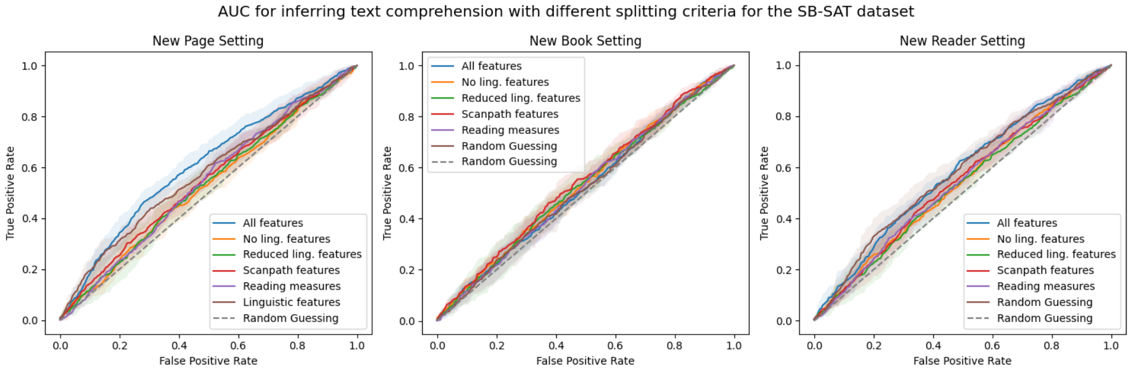


Figure 12: ROC curves for the SB-SAT dataset in the New Page, New Book and New Subject evaluation settings. For every experiment, the line represents the mean score of the different folds and the shaded area represents the standard error.

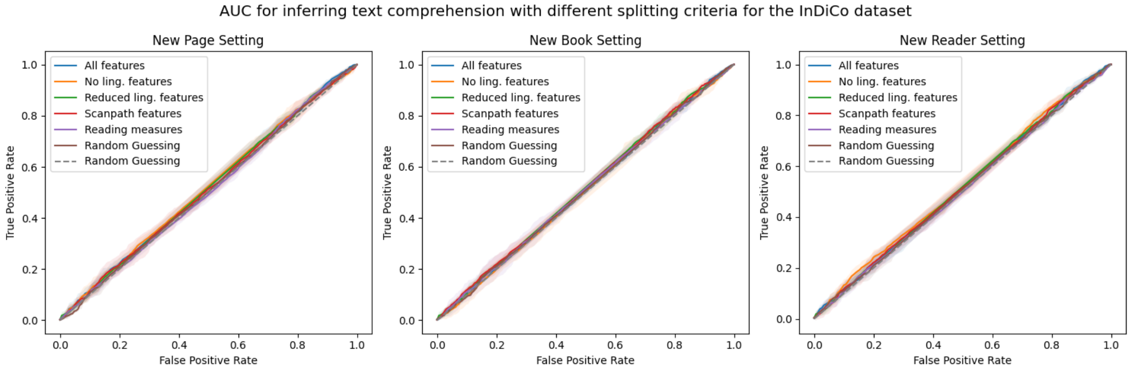


Figure 13: ROC curves for the InDiCo dataset in the New Page, New Book and New Subject evaluation settings. For every experiment, the line represents the mean score of the different folds and the shaded area represents the standard error.

in table 9. In the *New Page* setting, results are once again the best if all available features have been used as input. In the *New Book* setting, using only scanpath features lead to the best results again and the incorporation of linguistic features lead to noticeably worse results. The only difference in the results pattern between the datasets is in the *New Reader* setting where firstly, the model using only linguistic features was not on par with the model using all features and secondly, the best performance was achieved if linguistic features were omitted. This might once again be due to bad recording quality in entire sessions. All in all, it is difficult to derive a meaningful interpretation from the results in table 9 due to the data quality.

Features	**New Page**	**New Book**	**New Reader**
All features	**0.52 ± 0.018**	0.506 ± 0.012	0.515 ± 0.02
No ling. features	**0.52 ± 0.02**	0.504 ± 0.029	**0.524 ± 0.018**
Reduced ling. features	0.508 ± 0.019	0.509 ± 0.014	0.515 ± 0.007*
Scanpath features	0.513 ± 0.028	**0.516 ± 0.016**	0.514 ± 0.025
Reading measures	0.505 ± 0.016	0.509 ± 0.024	0.505 ± 0.017
Linguistic features	0.507 ± 0.014	0.507 ± 0.005	0.509 ± 0.017

Table 10: AUC results ± standard error for training LinguisticEyeLSTM on the InDiCo dataset with different features. The asterisk * indicates values that are significantly higher than random guessing. The colors indicate whether the score is higher (green) or lower (red) than the approach that used all linguistic features. Bold numbers highlight the best results in a given evaluation setting.

	All features	No ling. features	Reduced ling. features	Scanpath features	Reading measures	Linguistic features	Data type	Encoding
Vertical coordinate of the fixation	yes	yes	yes	yes	no	no	float	NA
Horizontal coordinate of the fixation	yes	yes	yes	yes	no	no	float	NA
Pupil size	yes	yes	yes	yes	no	no	float	NA
Current fixation duration	yes	yes	no	yes	no	no	int	NA
First fixation duration (FFD)	yes	yes	no	no	yes	no	int	NA
Total fixation duration (TFD)	yes	yes	no	no	yes	no	int	NA
Number of fixations	yes	yes	no	no	yes	no	int	NA
First-pass regression	yes	yes	no	no	yes	no	int	binary
Normalized Incoming Regression Count	yes	yes	no	no	yes	no	float	NA
Normalized Outgoing Progressive Saccade Count	yes	yes	no	no	yes	no	float	NA
Normalized Outgoing Regressive Saccade Count	yes	yes	no	no	yes	no	float	NA
Word length (number of characters)	yes	no	no	no	no	yes	int	NA
Sentence length (number of words)	yes	no	no	no	no	yes	int	NA
Mean word length within a sentence	yes	no	no	no	no	yes	float	NA
Text length (number of sentences)	yes	no	no	no	no	yes	int	NA
Simplified PoS-Tags	yes	no	yes	no	no	yes	object	OHE
Content word feature	yes	no	yes	no	no	yes	int	binary
Pronoun-noun ratio / Determiner-noun ratio	yes	no	no	no	no	yes	float	NA
Lexical density (sentence-level)	yes	no	no	no	no	yes	float	NA
Lexical density (text-level)	yes	no	no	no	no	yes	float	NA
Entity type	yes	no	no	no	no	yes	object	binary
Lexicalized Surprisal	yes	no	no	no	no	yes	float	NA
Lemma TTR	yes	no	no	no	no	yes	float	NA
Content lemma TTR	yes	no	no	no	no	yes	float	NA
Function lemma TTR	yes	no	no	no	no	yes	float	NA
More frequent synonym and/or homonym	yes	no	no	no	no	yes	int	binary
Lemma frequency	yes	no	no	no	no	yes	float	NA
Number of syntactic dependants to the right	yes	no	no	no	no	yes	int	NA
Number of syntactic dependants to the left	yes	no	no	no	no	yes	int	NA
Dependency distance to head	yes	no	no	no	no	yes	int	NA
Syntactic Surprisal	yes	no	no	no	no	yes	float	NA
Voice	yes	no	no	no	no	yes	float	NA
Semantic similarity between adjacent sentences	yes	no	no	no	no	yes	float	NA
Lexical overlap between adjacent sentences (lemma)	yes	no	no	no	no	yes	float	NA
TF-IDF	yes	no	no	no	no	yes	float	NA
Genre	yes	no	no	no	no	yes	object	OHE
Sentence is cut off	yes	no	no	no	no	yes	int	binary

Table 10: A list of features used as model input in the different experiments including the ablation study. Regarding the encoding column, OHE refers to one hot encoding, binary refers to binary features (0/1) and features that have the data type *float* or *int* don't require any further encoding.

8 Discussion and future work

In my approach to infer text comprehension from eye gaze data that was recorded during reading experiments, I framed the task as a binary classification problem and used scanpath features, reading measures and linguistic features of the stimulus text as inputs to a BiLSTM in order to infer whether the achieved score in text comprehension questions belonging to the text at hand were above or below the median for a given scanpath. I worked with the SB-SAT dataset which had been used for the same task in previous research and with a new corpus with German stimulus texts (InDiCo). My main contributions to this field are the annotation of the stimulus texts with new linguistic features based on findings in the field of psycholinguistics, the employment of a new neural network architecture for this task and the classification on a new dataset to provide baseline results. I evaluated my results in multiple cross-validation settings (*New Reader*, *New Page* and *New Book*). For the sake of comparability, my method has many similarities to the one presented in Reich et al. [2022], which hed the previous state of the art in this task, for instance the choice of scanpath features and reading measures and the evaluation strategy. Presently, there is only little research on the task of inferring text comprehension from eye gaze data and the results are not application ready yet. This leaves much room for creativity in how to approach this task but it is also limiting in that there are not many results available to compare my results with.

Generally, the results of my classification approach on the SB-SAT data were satisfactory, although not excellent. They exceeded the previous state of the art, especially in the *New Reader* evaluation setting. This is gratifying since Reich et al. [2022] expressed that generalizing to novel readers not seen during training was one of the main challenges for future work. In the other two evaluation settings, my approach was more or less on par with the state of the art for the SB-SAT dataset: I outperformed previous models by a very small margin but the standard deviation from the different folds suggests that my model isn't as good at adapting to new data as the previous state of the art model.

With regard to the research questions posed in chapter 1, it can be said that the linguistic annotation of the stimulus text increase the model performance in the task of inferring text comprehension from eye gaze data in the *New Page* and *New Reader* settings, where using only the linguistic features yielded almost the same results as using all available features. However, the use of linguistic features had a negative impact on the performance of the model in the *New Book* setting, suggesting that there is a trade-off between a model being able to generalize to new readers and pages and the model being able to generalize to new books.

Secondly, I investigated the impact of my network architecture. I compared the results for BEyeLSTM without GSFNet reported by Reich et al. [2022] with the results of my model using the same features as input. It appears that my network architecture falls short in comparison to BEyeLSTM without GSFNet, suggesting that a modular network consisting of multiple sequential subnets that take different features as inputs remains a promising approach to this problem.

Thirdly, the question of whether there a significant difference between the results of the classification of text comprehension between the two datasets can be answered with a clear *yes*. The performance on the SB-SAT surpassed chance level in the *New Page* and the *New Reader* settings but the performance on the InDiCo dataset didn't exceed chance level in any setting. The difference can be explained with the difference in data quality rather than with any other properties that set the datasets apart (such as language of the stimulus texts or distribution of the text across subjects).

For future work, the main goal will be to achieve application-relevant results. This could be attempted by adding even more linguistic features, since the introduction of additional linguistic features was beneficial in many cases. However, a middle ground should be found between the use of linguistic and scanpath features in order to account for the trade-off between the different evaluation settings. It would have been interesting to perform a more fine-grained ablation study to figure out which of the linguistic features were responsible for the drop of performance in the *New Book* setting but the time frame of this thesis would not allow for it. Additionally, it could be worthwhile to investigate whether using a pre-trained language model to extract contextualized embeddings of each word or paragraph improves the performance of the proposed architecture. Furthermore, regarding network architecture it would be interesting to experiment with a modular network architectures similar to BEyeLSTM without GSFNet consisting of three subnets which take reading measures, linguistic features and sequential features as inputs respectively. Then, the outputs of the three subnets could be weighted differently in order to attempt to find a balance between the features and lessen the trade-off between the different evaluation settings.

9 Conclusion

In this thesis, I presented LinguisticEyeLSTM, a sequential neural architecture that processes the sequence of fixations during reading and infers text comprehension. A key aspect of my approach is linguistically annotating the stimulus texts.
I was able to answer the following research questions:

- Does linguistic annotation of the stimulus texts improve the results in the task of inferring text comprehension from eye gaze data?
  Yes, as shown in section 7.2.2, the annotation of the stimulus texts with additional linguistic features proved to be beneficial in two of three evaluation settings (*New Page* and *New Reader*). However, there was a trade-off with the *New Book* setting which is generally the hardest to infer. In that setting, using linguistic features is a drawback.

- Does employing a single BiLSTM that takes all features as input yield better results than employing a network architecture consisting of several subnets?
  No, the model architecture I used did not outperform the modular architecture proposed in previous research, as discussed in section 7.2.2.

- Is there a difference between the results on the SB-SAT dataset and the InDiCo dataset?
  Yes, the performance on the SB-SAT dataset was much better than the performance on the InDiCo dataset (see section 7.1). However, I refrain from drawing a conclusions from this since I suspect the nature of the target label as well as the data quality of the latter dataset to be a potential issue.

My model outperforms the previous state of the art model which was a different architecture and included less linguistic features. It especially lead to an improvement in the ability to generalize to new readers, however generalizing to new books not seen during training remains a challenge. This thesis also confirms that findings from the field of psycholinguistics can be employed in order to improve the results of a neural model that takes eye gaze data as input. This is a small step towards models that can potentially one day reach application-relevant results, eliminating the need for laborious text comprehension assessments.

To conclude, my findings reinforce the notion that linguistic annotation of the stimulus texts is beneficial in tasks that infer labels that are related to cognitive processes from eye-gaze data using neural networks. Future research should focus on developing application-relevant models by further exploring model architectures and input features with an emphasis on alleviating the trade-off between different evaluation setting occurring due to the use of scanpath and linguistic features.

References

S. Ahn, C. Kelton, A. Balasubramanian, and G. Zelinsky. Towards Predicting Reading Comprehension From Gaze Behavior. In *ACM Symposium on Eye Tracking Research and Applications*, pages 1–5, 2020.

H. Al-Samarraie, A. Eldenfria, and H. Dawoud. The impact of personality traits on users' information-seeking behavior. In *Information Processing Management*, pages 237–247, 2017.

S. Arehalli, B. Dillon, and T. Linzen. Syntactic Surprisal From Neural Models Predicts, But Underestimates, Human Processing Difficulty From Syntactic Ambiguities. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 301–313, 2022.

O. Augereau, H. Fujiyoshi, and K. Kise. Towards an automated estimation of english skill via toeic score based on reading analysis. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 1285–1290, 2016.

J. J. S. Barton, H. M. Hanif, L. E. Björnström, and C. Hills. The word-length effect in reading: A review. *Cognitive Neuropsychology*, 31(5-6):378–412, 2014.

Y. Berzak, C. Nakamura, S. Flynn, and B. Katz. Predicting native language from gaze. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 541–551, 2017.

Y. Berzak, B. Katz, and R. Levy. Assessing Language Proficiency from Eye Movements in Reading. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1986–1996, 2018.

J. F. Boisvert and N. D. Bruce. Predicting task from eye movements: On the importance of spatial distribution, dynamics, and image features. *Neurocomputing*, 207:653–668, 2016.

D. Brimo, E. Lund, and A. Sapp. Syntax and reading comprehension: a meta-analysis of different spoken-syntax assessments. *International Journal of Language & Communication Disorders*, 53(3):431–445, 2018.

L. Copeland and T. Gedeon. Measuring reading comprehension using eye movements. In *2013 IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom)*, pages 791–796, 2013.

S. H. Deacon and K. Francis. The relationship of morphological analysis and morphological decoding to reading comprehension. *Journal of Research in Reading*, 40(1):1–16, 2017.

L. C. Ehri. *Word Recognition in Beginning Literacy*, chapter Grapheme–Phoneme Knowledge Is Essential for Learning to Read Words in English, pages 3–40. Routledge, New York, 1 edition, 1998.

C. Fellbaum. Wordnet. an electronic lexical database, 1998. Accessed on 25.05.2023.

S. Frank and R. Thompson. Early effects of word surprisal on pupil size during reading. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, pages 1554–1559, 2012.

L. S. Fuchs, D. Fuchs, and L. Maxwell. The validity of informal reading comprehension measures. *Remedial and Special Education*, 9(2):20–28, 1988.

M. A. Gernsbacher and M. P. Kaschak. Text compehension. In D. Reisberg, editor, *The Oxford Handbook of Cognitive Psychology*, chapter 29, pages 462–474. Oxford University Press, Oxford, 2013.

J. Hale. A probabilistic earley parser as a psycholinguistic model. In *North American Chapter of the Association for Computational Linguistics*, pages 1–8, 2001.

B. Hamp and H. Feldweg. Germanet - a lexical-semantic net for german. In *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, 1997.

N. Harkio and P. Pietilä. The role of vocabulary breadth and depth in reading comprehension: A quantitative study of finnish efl learners. *Journal of Language Teaching and Research*, 7(6):1079–1088, 2016.

J. M. Henderson, S. V. Shinkareva, J. Wang, S. G. Luke, and J. Olejarczyk. Predicting cognitive state from eye movements. *PLOS ONEE*, 8(5):1–6, 2013.

V. Henrich and E. Hinrichs. Gernedit - the germanet editing tool. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*, pages 2228–2235, 2010.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

N. Hollenstein, M. Tröndle, M. Plomecka, S. Kiegeland, Y. Özyurt, L. A. Jäger, and N. Langer. Reading task classification using eeg and eye-tracking data, 2021.

M. Honnibal and I. Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

S. Hoppe, T. Loetscher, S. A. Morey, and A. Bulling. Eye movements during everyday behavior predict personality traits. *Frontiers in Human Neuroscience*, 12, 2018.

M. Iqbal, M. Noor, F. Muhabat, and B. Kazemian. Factors responsible for poor english reading comprehension at secondary level. *Communication Linguistics Studies*, 1:1–6, 05 2015.

M. Jaan. Parts of speech in predicting reading comprehension. *Journal of Quantitative Linguistics*, 4(1-3):156–163, 1997.

T. Jagaiah, N. G. Olinghouse, and D. M. Kearns. Syntactic complexity measures: variation by genre, grade-level, students' writing abilities, and writing quality. *Reading and Writing*, pages 1–62, 2020.

L. A. Jäger, S. Makowski, P. Prasse, S. Liehr, M. Seidler, and T. Scheffer. Deep eyedentification: Biometric identification using micro-movements of the eye. In *Machine Learning and Knowledge Discovery in Databases; European Converence, ECML PKDD 2019*, pages 299–314, 2019.

M. A. Just and P. A. Carpenter. A theory of reading: from eye fixations to comprehension. *Psychological Review*, 87(4):329–354, 1980.

B. B. Kadayat and E. Eika. Impact of sentence length on the readability of web for screen reader users. In *Universal Access in Human-Computer Interaction. Design Approaches and Supporting Technologies: 14th International Conference, UAHCI 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part I*, page 261–271, Berlin, Heidelberg, 2020. Springer-Verlag.

M. J. Kieffer and N. K. Lesaux. Breaking down words to build meaning: Morphology, vocabulary, and reading comprehension in the urban classroom. *The Reading Teacher*, 61(2):134–144, oct 2007.

C. Klein-Braley. A cloze-up on the c-test: A study in the construct validation of authentic tests. *Language Testing*, 2:76–104, 1985.

H. Kolb. *Gross Anatomy of the Eye.* University of Utah Health Sciences Center, Salt Lake City (UT), 1995.

K. Kunze, Y. Utsumi, Y. Shiga, K. Kise, and A. Bulling. I know what you are reading: recognition of document types using mobile eye tracking. In *ISWC 2013 - Proceedings of the 2013 ACM International Symposium on Wearable Computers*, pages 113–116, 09 2013.

R. Levy. Expectation-based syntactic comprehension. *Cognition*, 106(3): 1126–1177, 2008.

D. J. Lohr, H. K. Griffith, S. Aziz, and O. V. Komogortsev. A metric learning approach to eye movement biometrics. *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–7, 2020.

Y. Lou, Y. Liu, J. Kaakinen, and X. Li. Using support vector machines to identify literacy skills: Evidence from eye movements. *Behavior Research Methods*, 49: 887–895, 2016.

S. Makowski, L. A. Jäger, A. Abdelwahab, N. Landwehr, and T. Scheffer. A discriminative model for identifying readers and assessing text comprehension from eye movements, 2019.

S. Makowski, P. Prasse, D. R. Reich, D. Krakowczyk, L. A. Jäger, and T. Scheffer. DeepEyedentificationLive: Oculomotoric Biometric Identification and Presentation-Attack Detection Using Deep Neural Networks. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(4):506–518, Oct. 2021.

C. B. Marks, M. J. Doctorow, and M. C. Wittrock. Word frequency and reading comprehensiony. *The Journal of Educational Research*, 67(6):259–262, 1974.

P. Martínez-Gómez and A. Aizawa. Recognition of understanding level and language skill using measurements of reading behavior. In *Proceedings of the 19th International Conference on Intelligent User Interfaces*, IUI '14, page 95–104, New York, NY, USA, 2014.

S. H. Moghadam, Z. Zainal, and M. Ghaderpour. A review on the important role of vocabulary knowledge in reading comprehension performance. *Procedia - Social and Behavioral Sciences*, 66:555–563, 2012.

D. C. Mézière, L. Yu, E. Reichle, T. von der Malsburg, and G. McArthur. Using eye-tracking measures to predict reading comprehension, Sep 2021.

R. L. Nurjanah. The analysis on students' difficulties in doing reading comprehension final test. *Methathesis: Journal of English language, literature and teaching*, 2(2):253–264, 2018.

J. J. Pikulski and D. J. Chard. Fluency: Bridge between decoding and reading comprehension. *The Reading Teacher*, 58(6):510–519, 2005.

G. S. Pinnell, J. J. Pikulski, K. K. Wixson, J. R. Campbell, P. B. Gough, and A. S. Beatty. *Listening to Children Read Aloud. Data from NAEP's Integrated Reading Performance Record (IRPR) at Grade 4*. Washington, DC: Office of Educational Research and Improvement, U.S. Department of Education, 1995.

P. Raatikainen, J. Hautala, O. Loberg, T. KÃ€rkkÃ€inen, P. LeppÃ€nen, and P. Nieminen. Detection of developmental dyslexia with machine learning using eye movement data. *Array*, 12:Article 100087, 2021.

K. Rayner. The 35th sir frederick bartlett lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, 62(8):1457–1506, 2009. doi: 10.1080/17470210902816461. URL `https://doi.org/10.1080/17470210902816461`. PMID: 19449261.

K. Rayner, K. Chace, T. Slattery, and J. Ashby. Eye movements as reflections of comprehension processes in reading. *Scientific Studies of Reading - SCI STUD READ*, 10:241–255, 07 2006. doi: 10.1207/s1532799xssr1003_3.

D. R. Reich, P. Prasse, C. Tschirner, P. Haller, F. Goldhammer, and L. A. Jäger. Inferring Native and Non-Native Human Reading Comprehension and Subjective Text Difficulty from Scanpaths in Reading. In *2022 Symposium on Eye Tracking Research and Applications*, ETRA '22, pages 1–8, New York, NY, USA, June 2022. Association for Computing Machinery.

E. Reichle, A. Reineberg, and J. Schooler. Eye movements during mindless reading. *Psychological science*, 21:1300–10, 09 2010.

I. R. Rodríguez-Ortiz, F. J. Moreno-Pérez, I. C. Simpson, M. Valdés-Coronel, and D. Saldaña. The influence of syntactic knowledge on reading comprehension varies as a function of oral vocabulary in spanish-speaking children. *Journal of Research in Reading*, 44(3):695–714, 2021.

M. Schijndel and T. Linzen. Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, 45, 06 2021.

E. R. Schotter, B. Angele, and K. Rayner. Parafoveal processing in reading. *Atten Percept Psychophys*, 74(1):5–35, 2012.

M. Shojaeizadeh, S. Djamasbi, R. C. Paffenroth, and A. C. Trapp. Detecting task demand via an eye tracking machine learning system. *Decision Support Systems*, 116:91–101, 2019.

C. Snow. *Reading for Understanding:: Toward an RD Program in Reading Comprehension.* RAND, Pittsburgh, 2002.

R. Speer. rspeer/wordfreq: v3.0, Sept. 2022.

K. E. Stanovich. Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, 21: 360–407, 1986.

S. Stuart, editor. *Eye Tracking. Background, Methods, and Applications.* Humana New York, NY, 2022.

P. G. TestDaF. Testdaf: Konzeption, stand der entwicklung, perspektiven. *Zeitschrift für Fremdsprachenforschung*, 11(1):63–82, 2000.

G. Underwood, A. Hubbard, and H. Wilkinson. Eye fixations predict reading comprehension: The relationships between reading skill, reading speed, and visual inspection. *Language and Speech*, 33(1):69–81, 1990.

R. Zipoli. Unraveling difficult sentences: Strategies to support reading comprehension. *Intervention in School and Clinic*, 52, 08 2016.

A Tables

subject	session	text	screen	CURRENT_FIX_INTEREST_AREA				
				X	Y	ID	LABEL	RUN_ID
4	2	1	1	139.4	136.2	2	o	1
4	2	1	1	203.1	135.1	7	w	1
4	2	1	1	240.4	138.8	10	d	1
4	2	1	1	396.2	136.7	22	l	1
4	2	1	1	523.3	131.3	32	c	1
4	2	1	1	667.5	173.4	83	_	2
4	2	1	1	761.7	154.4	83	_	2
4	2	1	1	862.2	136.8	58	g	1
4	2	1	1	691.5	145.5	83	_	3
4	2	1	1	640.3	145.4	83	_	3
4	2	1	1	706.3	151.5	83	_	3
4	2	1	1	875.4	143.5	83	_	3
4	2	1	1	1050.8	126.1	72	M	1
4	2	1	1	1038.0	124.2	71	_	1
4	2	1	1	1141.0	138.1	79	n	1
4	2	1	1	1190.2	144.9	83	_	4
4	2	1	1	245.5	188.8	93	o	1

Table 11: InDiCo fixation report for participant 4, session 2, text 1, screen 1 after data cleaning. There is one fixation per line, with the vertical and horizontal location of the fixation. Interest areas are on the level of character after exporting the fixation report from SR Research Dataviewer. The interest areas are indexed and the variable RUN_ID denotes how many times the current character has already been fixated. Underscores signify fixations outside an interest area.

subject	session	text	screen	CURRENT_FIX_INTEREST_AREA				
				X	Y	ID	LABEL	RUN_ID
4	2	1	1	139.4	136.2	1	Wohin	1
4	2	1	1	203.1	135.1	2	wird	1
4	2	1	1	240.4	138.8	2	wird	2
4	2	1	1	396.2	136.7	5	Klima	1
4	2	1	1	523.3	131.3	6	entwickeln?	1
4	2	1	1	862.2	136.8	10	Folgen	1
4	2	1	1	1050.8	126.1	13	Menschen	1
4	2	1	1	1141.0	138.1	13	Menschen	2
4	2	1	1	245.5	188.8	15	Mitteleuropa	1
4	2	1	1	182.0	190.4	15	Mitteleuropa	2
4	2	1	1	391.8	180.6	17	rechnen?	1
4	2	1	1	1144.4	185.7	26	hin	1
4	2	1	1	1191.3	188.7	27	zu	1
4	2	1	1	245.5	225.3	28	Befürchtungen,	1
4	2	1	1	356.4	225.2	29	dass	1
4	2	1	1	467.8	229.5	31	wegen	1
4	2	1	1	537.1	237.4	32	des	1

Table 12: InDiCo fixation report for participant 4, session 2, text 1, screen 1 after fixation mapping. There is one fixation per line, with the vertical and horizontal location of the fixation. Interest areas are on the level of words. The interest areas are indexed and the variable RUN_ID denotes how many times the current character has already been fixated. Underscores signify fixations outside an interest area.

X	Y	ID	LABEL	RUN_ID	ffd	tfd	n_fix	fpr	n_regressions_norm
139.4	136.2	1	Wohin	1	256	256	1	0	0.0
203.1	135.1	2	wird	1	192	370	2	0	0.0
240.4	138.8	2	wird	2	192	370	2	0	0.0
396.2	136.7	5	Klima	1	128	128	1	0	0.0
523.3	131.3	6	entwickeln?	1	313	313	1	0	0.0
862.2	136.8	10	Folgen	1	138	138	1	0	0.0
1050.8	126.1	13	Menschen	1	336	506	2	0	0.0
1141.0	138.1	13	Menschen	2	336	506	2	0	0.0
245.5	188.8	15	Mitteleuropa	1	147	332	2	0	0.018182
...	...	...	...	...	...	...	...	...	...
980.9	541.7	99	Patzelt	1	269	437	2	0	0.0
1001.0	540.8	99	Patzelt	2	269	437	2	0	0.0
162.3	589.7	102	3000	1	267	267	1	0	0.009091
259.6	578.2	103	Quadratkilometer	1	166	166	1	0	0.0
603.5	585.7	107	dieser	1	304	304	1	0	0.0
257.5	633.2	113	Gleichzeitig	1	110	431	3	0	0.018182
191.1	626.1	113	Gleichzeitig	2	110	431	3	0	0.018182
232.0	635.2	113	Gleichzeitig	3	110	431	3	0	0.018182
435.0	628.4	116	aufgrund	1	229	229	1	0	0.0
644.8	632.5	118	wärmeren	1	243	243	1	0	0.0
771.3	635.3	119	Temperaturen	1	231	231	1	0	0.0
292.1	677.2	127	verschoben.	1	153	292	2	0	0.009091
222.6	683.1	126	oben	1	173	173	1	1	0.0
333.2	675.3	127	verschoben.	2	153	292	2	0	0.009091

Table 13: InDiCo reading measures for participant 4, session 2, text 1, screen 1. There is one fixation per line with interest areas on the token-level and fixations outside of interest areas removed. ID and label indicate which word has been fixated, X and Y denote the horizontal and vertical position of the fixation and Run ID is the running count of the number of fixations on a word. For a description of the reading measures, see section 5.1.2.2

X	Y	ID	LABEL	RUN_ID	ffd	tfd	n_fix	fpr	n_regressions_norm
348.7	182.0	21	long	1	26	364	2	0	0.0
630.5	400.3	69	safe	1	216	364	2	1	0.006849
492.0	400.2	66	boundless	1	125	315	2	0	0.006849
526.6	390.5	67	world,	1	486	1040	2	0	0.0
545.8	397.9	67	world,	1	486	1040	2	1	0.0
525.7	181.8	24	other	1	158	372	2	1	0.013699
167.3	124.8	1	Even	1	194	194	1	0	0.006849
270.9	111.3	3	my	1	210	449	2	0	0.006849
344.1	115.3	5	friends	1	228	228	1	1	0.0
242.6	113.3	3	my	2	210	449	2	0	0.006849
225.6	114.6	2	then	1	117	117	1	0	0.0
327.4	119.3	4	only	1	174	174	1	0	0.0
476.0	115.4	7	made	1	261	261	1	0	0.0
604.5	126.7	9	paper	1	255	255	1	0	0.0
691.2	128.0	11	ink.	1	179	179	1	0	0.0
794.7	123.7	13	school	1	135	135	1	0	0.0
854.8	127.6	15	had	1	149	149	1	0	0.0
92.0	207.0	16	learned	1	186	186	1	0	0.006849
...	...	...	...	...	...	...	...	...	...

Table 14: SB-SAT reading measures for participant 1, session 1, text 1, screen 1. There is one fixation per line with interest areas on the token-level and fixations outside of interest areas removed. ID and label indicate which word has been fixated, X and Y denote the horizontal and vertical position of the fixation and Run ID is the running count of the number of fixations on a word. For a description of the reading measures, see section 5.1.2.2

Dataset	Text ID	Text Name	Topic	Genre
InDiCo	1	Auf dem Weg zu einer Eiszeit. Wie die Akademie der Wissenschaften die Zukunft sieht	Article about climate change and the impact of human activity on the climate.	scientific
InDiCo	2	Forschungsprojekt Eltern-Kind-Interaktion. Spielen Väter mit ihren Kindern anders als Mütter?	Article about a study on the difference between father-child interaction and mother-child interaction.	scientific
InDiCo	3	Nüsse: Harte Schale, gesunder Kern	Article about the origin and the nutritional benefits of nuts, as well as their impact on health and potential risks associated with their consumption.	informative
InDiCo	4	Die Sammelleidenschaft beim Menschen	Article about the human tendency to collect things.	informative
InDiCo	5	Lernen: Lust oder Last?	Article abouth the concept of a learning drive in humans ad how the brain responds to learning.	informative
InDiCo	6	Zur Funktionsweise von Werbung	Article about the psychological aspects of advertising.	informative
InDiCo	7	Unheimlich, aber normal. Haben die letzten Erdbeben eine gemeinsame Ursache?	Article about the causes of earthquakes.	informative
InDiCo	8	Wie wir riechen und was wir riechen	Article about a study on the sense of smell in monkeys and humans.	scientific
InDiCo	9	Studium per Mausklick	Article about the concept, the advantages and challenges of virtual universities in Germany, Canada and the UK and the USA.	informative
InDiCo	10	Die biologische Uhr	Article about the concept of chronobiology, the adaptation of organisms to the timing of their environment.	scientific
InDiCo	11	Kommunikation mit Pflanzen	Article about the topic of plant communication (effects of words and sounds on pant growth).	informative

Table 15: Stimulus texts of the InDiCo Dataset.

Dataset	Text ID	Text Name	Topic	Genre
SB-SAT	1	Dickens	A memoir of a young boy who finds solace through reading despite his fathers wishes. He finds refuge in a local bookshop and recieves Charles Dickens' "Great Expectations" form the shop owner. The book inspires the boy to become an author.	fiction
SB-SAT	2	Flytrap	An article about the Venus flytrap's ability to detect and capture prey.	scientific
SB-SAT	3	Genome	An article about genetic engineering and the use of transgenic animals to produce pharmaceuticals.	scientific
SB-SAT	4	Northpole	A memoir describing the emotions through a narator who is embaking on a dangerous journey to the Nort pole.	fiction

Table 16: Stimulus texts of the SB-SAT dataset.