DISS. ETH NO. 27260

Leveraging Cognitive Processing Signals for Natural Language Understanding

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

NORA HOLLENSTEIN

MSc in Artificial Intelligence, University of Edinburgh

born on 30.09.1990

citizen of Mosnang, SG, Switzerland

accepted on the recommendation of

Prof. Dr. Ce Zhang (ETH Zurich), examiner
Prof. Dr. Joachim Buhmann (ETH Zurich), co-examiner
Prof. Dr. Nicolas Langer (University of Zurich), co-examiner
Prof. Dr. Lisa Beinborn (Vrije Universiteit Amsterdam), co-examiner
Prof. Dr. Martin Volk (University of Zurich), co-examiner

2021

DS3Lab
Institute for Computing Platforms
ETH Department of Computer Science

NORA HOLLENSTEIN

A dissertation submitted to
ETH Zurich
for the degree of Doctor of Sciences

DISS. ETH NO. 27260

examiner:
Prof. Dr. Ce Zhang
co-examiners:
Prof. Dr. Joachim Buhmann
Prof. Dr. Nicolas Langer
Prof. Dr. Lisa Beinborn
Prof. Dr. Martin Volk

Examination date: January 28, 2021

Leveraging Cognitive Processing Signals for Natural Language Understanding

Abstract

In this thesis, we aim to narrow the gap between human language processing and computational language processing. Natural language processing (NLP) models are imperfect and lack intricate capabilities that humans access automatically when processing speech or reading text. Human language processing signals can be leveraged to increase the performance of machine learning (ML) models and to pursue explanatory research for a better understanding of the differences between human and machine language processing. In particular, the contributions of this thesis are threefold:

1. We compile the *Zurich Cognitive Language Processing Corpus (ZuCo)*, a dataset of simultaneous eye tracking and electroencephalography (EEG) recordings from participants reading natural sentences from real-world texts. When we read, our brain processes language and generates cognitive processing signals such as gaze patterns and brain activity. ZuCo includes data of 30 English native speakers, each reading 700-1,100 sentences. This corpus represents a valuable resource for cognitively-inspired NLP.

2. We leverage these cognitive signals to augment ML models for NLP. Compared to purely text-based models, we show consistent improvements across a range of tasks and for both eye tracking and brain activity data. We further explore two of the main challenges in this area: (i) decoding brain activity for language processing and (ii) dealing with limited training data to eliminate the need for recorded cognitive signals at test time.

3. We evaluate the cognitive plausibility of computational language models, the cornerstones of state-of-the-art NLP. We develop CogniVal, the first openly available framework for evaluating English word embeddings based on cognitive lexical semantics. Specifically, embeddings are evaluated by their performance at predicting a wide range of cognitive data sources recorded during language comprehension, including multiple eye tracking datasets and brain activity recordings such as electroencephalography and functional magnetic resonance imaging.

ZUSAMMENFASSUNG

Diese Arbeit bezweckt, eine Brücke zwischen menschlicher Sprachverarbeitung und maschineller Sprachverarbeitung (Natural Language Processing; NLP) zu schlagen. Machine Learning (ML) Modelle für die maschinelle Sprachverarbeitung sind fehlerhaft und verfügen nicht über die komplexen Fähigkeiten, auf die Menschen automatisch zugreifen, wenn sie gesprochene Sprache verarbeiten oder einen Text lesen. Signale der kognitiven menschlichen Verarbeitung von Sprache können genutzt werden, um die Leistung von NLP Modellen zu steigern. Dies fördert auch das Verständnis der Unterschiede zwischen der menschlichen und der maschinellen Sprachverarbeitung. Die Beiträge dieser Arbeit sind in den folgenden drei Punkten zusammengefasst:

1. Wir haben das *Zurich Cognitive Language Processing Corpus (ZuCo)* erstellt, einen Datensatz mit simultanen Eye-Tracking- und Elektroenzephalographie-Aufzeichnungen (EEG) von Probanden, die natürliche Sätze lesen. Beim Lesen verarbeitet unser Gehirn die Sprache und generiert kognitive Signale wie Blickmuster und Gehirnaktivität, die mit verschiedenen Methoden aufgezeichnet werden können. ZuCo enthält Daten von 30 englischen Muttersprachler/innen, die jeweils 700 bis 1100 Sätze lesen. Dieses Korpus ist eine wertvolle Ressource für kognitiv inspirierte maschinelle Sprachverarbeitung.

2. Wir entwickeln Modelle für das maschinelle Lernen von NLP Anwendungen, die diese kognitiven Verarbeitungssignale als zusätzliche Informationsquelle nutzen. Wir zeigen konsistente Verbesserungen gegenüber text-basierten Modellen für eine Reihe von Sprachverarbeitungsaufgaben sowohl mit Eye-Tracking- als auch mit Gehirnaktivitätsdaten. Wir untersuchen zusätzlich zwei der Hauptherausforderungen, auf die wir in diesem Bereich gestossen sind: (i) das Dekodieren von Gehirnaktivitätsdaten für die maschinelle Sprachverarbeitung und (ii) den Umgang mit begrenzten Trainingsdaten, damit die Notwendigkeit von aufgezeichneten kognitiven Signalen zur Testzeit entfällt.

3. Wir werten die kognitive Plausibilität von maschinellen Sprachmodellen aus, den Eckpfeilern der modernen automatischen Sprachverarbeitung. Basierend auf der Theorie der

kognitiven lexikalischen Semantik haben wir CogniVal entwickelt, das erste öffentlich verfügbare Framework zur Bewertung von englischen Wortrepräsentationen. Insbesondere werden Wortrepräsentationen aufgrund ihrer Leistung bei der Vorhersage einer Vielzahl kognitiver Datenquellen bewertet, einschließlich mehrerer Datensätze von Augenbewegungen und Hirnaktivität, die während des Sprachverständnisses aufgezeichnet wurden.

Acknowledgments

First of all, I wish to express my sincere appreciation to Ce Zhang for being a great advisor and for guiding me through the world of academia. I learned so many things from him that have helped me become a better researcher. His persistent help was invaluable for realizing my projects. The support and scientific freedom he provided within the lively and honest research environment at the ETH Systems Group were decisive to the outcome of this thesis.

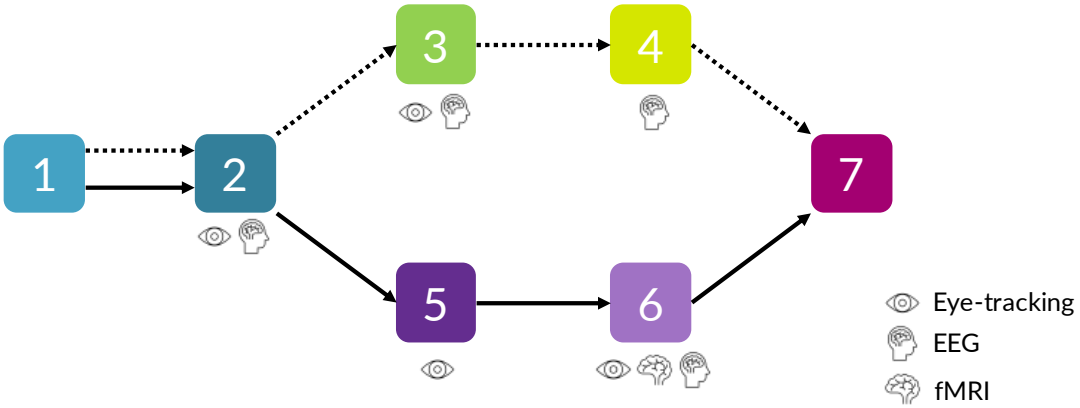
I thank the members of my committee for taking the time to assess this thesis, and more importantly, for their contributions to my work and career. I am grateful to my co-advisor Joachim Buhmann for the helpful discussions, to Nicolas Langer for his enormous help with the data collection and all follow-up projects, to Lisa Beinborn for the fruitful collaborations, and to Martin Volk for the opportunities and feedback he has provided me ever since my first NLP course.

I am grateful to my parents, my brothers, and my friends for distracting me when I needed it and for putting my research endeavours into perspective. Gracias por el apoyo!
Most importantly, thank you Simon for making everything brighter.

And last but not least, I am thankful to the members of the morning council for the countless coffee meetings, the technical discussions, the after-work swimming and beer sessions, and for reading this thesis. Thank you for the fun times during this PhD adventure.

Zurich, January 2021

Reading Threads



The figure above shows the possible reading threads of this thesis, split by topics of interest: If the reader is interested in augmenting and improving natural language understanding models with cognitive processing signals, then follow the chapters according to the dashed arrows. If the reader is interested in leveraging human language processing signals for the evaluation and interpretability of NLP models, then follow the solid arrows. Otherwise, you can read this thesis in linear order. Additionally, each chapter is marked with symbols showing which type of cognitive processing signals were included.

Contents

Chapter 1

Introduction

In the first chapter of this dissertation, we introduce the research area of cognitively-inspired natural language processing and motivate this line of work. Furthermore, we state the research questions to be addressed in the scope of this thesis and we present the contributions and impact of the presented work.

1.1 Cognitively-Inspired Natural Language Processing

Cognitively-inspired natural language processing (NLP) is a research field that has received increased interest in the past few years. It aims to develop NLP applications which are inspired and guided by the human language processing mechanisms. On the one hand, cognitively-inspired NLP uses cognitive processing signals to advance computational language processing. On the other hand, it enables exploratory research to further our understanding of the human language process using computational modeling techniques.

Supervised machine learning (ML) algorithms depend on large amounts of training samples. However, human language processing data comes as a highly heterogeneous mix of linguistic, behavioral, and physiological datasets (e.g., electrical brain activity, neuroimaging, eye movements, electrodermal activity, as well as behavioral norms) of limited size (Blache et al., 2018; Futrell et al., 2018; Lopopolo et al., 2018). Hence, extracting the linguistic structure and eliminating noise from other parallel cognitive processes is the first challenge in cognitively-inspired NLP (Minnema and Herbelot, 2019; Hollenstein et al., 2020a). During this thesis, more datasets

of human language processing signals from neuroscience and psychology have been made available, including our Zurich Cognitive Language Processing Corpus, and are shifting their focus from controlled constructed linguistic stimuli to more naturalistic experiment paradigms of processing real-world language stimuli (Hamilton and Huth, 2018). Therefore, such cognitively-inspired datasets are becoming more practical and interesting for the field of computational linguistics (Artemova et al., 2020). In this thesis, we leverage a wide range of these neurolinguistic datasets, noting that it is a challenge to learn suitable representations of these data and to combine multiple modalities in ML models.

Recent NLP research has demonstrated that the incorporation of behavioural and cognitive data, mostly in the form of eye movement signals, can improve the modelling quality for a variety of complex NLP tasks, for instance, semantic tasks such as sentiment analysis (Long et al., 2017) and sarcasm detection (Mishra et al., 2016b), but also syntactic tasks such as part-of-speech tagging (Barrett et al., 2016). However, it remains to be empirically analyzed whether these improvements generalize across tasks and across cognitive data types.

Data sources of human language processing signals also provide a robust benchmark for the evaluation of essential components of several NLP systems (Bakarov, 2018b). The quality of computational language representations can be assessed with cognitive language processing data. It has been shown that semantic representations can be decoded successfully from brain imaging data (Huth et al., 2016). These approaches mostly focus on one modality of brain activity data from small individual cognitive datasets (e.g., Abnar et al. (2018) and Rodrigues et al. (2018)). The small number and the size of data sources have limited their use for the evaluation of word embeddings until now (Bakarov, 2018b). While neural language models have become increasingly popular, our understanding of these black box algorithms is still rather limited (Gilpin et al., 2018). For a truly intrinsic evaluation of language models more research about the cognitive plausibility of current language models is required (Ettinger, 2020; Manning et al., 2020). Previous work has shown that state-of-the-art English language models accurately predict language processing in the brain (Schrimpf et al., 2020; Merkx and Frank, 2020). Hence, there are still many open research questions to be tackled at the intersection of interpretability of NLP models and cognitive science.

Within the scope of this thesis, we address some of the challenges described above and contribute to the field of cognitively-inspired natural language processing. We aim to reduce the gap between human language processing and computational language processing. In the following section, we pose the three main research questions to be tackled in this thesis.

1.2 Motivation

The human brain processes many tasks unconsciously and without visible effort. One such task is language processing. Someone speaks, the sound enters our ears and we understand its meaning within milliseconds. Similarly, when we see a string of characters, we can almost immediately distinguish whether it is nonsense or a meaningful sentence; and once we have read it, we grasp not only the meaning but also the syntactic structure of this sentence. These processes are complex computational problems and our brains contain dedicated information processing machinery to solve these tasks. Natural questions one might ask are whether we can capture this mental representation of language and use it to improve our machine learning systems; or whether computational models and cognitive signals can be aligned to obtain more interpretable NLP models?

To this end, we aim to find and extract relevant aspects of text understanding and annotation directly from the source, i.e., eye tracking and brain activity signals during reading. We leverage passive cognitive data from humans processing language, with techniques such as eye tracking and brain activity measurements. The main challenge in working with this type of data is its availability and its noisiness. We decode the signals so that we can then efficiently use them to improve and evaluate machine learning systems for NLP.

The main goal of this work is to explore the relationship between human intelligence and machine intelligence. More specifically, how our understanding of human language processing can benefit machine learning based natural language processing, and vice versa. In order to approach this broad research topic, the following three research questions have been defined as a guiding thread of concrete, feasible and testable milestones for the projects completed in this dissertation.

1. Can we compile a dataset of recorded human language processing signals which fulfills the state of the art in neuroscience and is usable for ML applications?
Since many datasets in neuroscience are not freely available and have been developed under strict experimental conditions, we build a dataset tailored specifically to these research topics. We collect cognitive language processing signals, i.e., eye tracking and electroencephalography recordings during various reading tasks. This data has three main characteristics that distinguish it from most previous neurolinguistic datasets: (1) the dataset is openly available to allow faster advances in this field, (2) it records the reading process as naturally as possible, and (3) it contains two types of cognitive signals for co-registration analyses.

2. Can signals recorded during human language processing be applied to improve machine learning based NLP tasks?

We then use these brain activity and eye tracking signals recorded during human language processing to improve machine learning based NLP models for various NLP tasks, including named entity recognition and sentiment analysis. To this end, we analyze this type of data and develop multi-modal machine learning methods. It has been shown that certain NLP tasks can be improved with eye tracking data. However, it has not been shown yet whether these findings generalize across cognitive data types and across NLP tasks.

We further analyze two of the main challenges we encountered. First, we explore methods of decoding electrical brain activity signals for language processing. Using brain activity data for this type of application is a very recent approach. Thus, there is a need to discover and define best practices in how to leverage human data for machine learning. This involves both cognitively plausible preprocessing steps as well as data-driven feature extraction methods. We assess challenges such as noise in the data and variance between subjects, since it is essential to understand and differentiate the signal from the noise in brain activity data.

Second, we address the challenge of learning from limited data. Since neurolinguistic datasets are expensive to record, their size is often limited. Therefore, the number of samples we can extract for machine learning applications is relatively small. Moreover, while we might have enough data to train these models on, we do not want to require cognitive data at test time in order to enhance the applicability of these augmented models. To counteract this problem, we analyze certain strategies of aggregating and predicting cognitive features, so that no real-time recordings are required during inference.

3. Can human language processing signals be applied to evaluate the quality and cognitive plausibility of computational language models?

Pre-trained contextualized language models have become the cornerstones of state-of-the-art NLP models. Deep contextualized word representations model both complex syntactic and semantic characteristics of word use, and how these uses vary across linguistic contexts. Unfortunately, current state-of-the-art machine learning algorithms for language understanding are still mostly black box algorithms. The link between a vector of numbers and a humanly interpretable representation of semantics is still hidden. This means that we cannot comprehend or track the decision-making process of NLP models. However, interpretability is the key for many NLP applications to be able to understand the algorithms' decisions. Moreover, one of the challenges

for computational linguistics is to build cognitively plausible models of language processing, i.e., models that integrate multiple aspects of human language processing at the syntactic and semantic level.

Evaluating and comparing the quality of different word representations is a well-known, largely open challenge. In this thesis, we build the first framework for cognitive word embedding evaluation. We evaluate word representations based on cognitive lexical semantics, i.e., by how much they reflect the semantic representations in the human brain.

1.3 Contributions

First, we collected exactly the type of data required for the proposed research avenue. The Zurich Cognitive Language Processing Corpus is openly available to the research community and tailored specifically to cognitively-inspired NLP research questions. The co-registration of eye movements and brain activity allows for linguistic analyses on multiple levels, from lexical processing to full sentence processing. We also provide a growing collection of available neurolinguistic data sources for cognitively-inspired NLP.[1] Furthermore, our research has provided a better understanding of the use of brain activity signals in language processing and for machine learning for various communities (both computer science and neuroscience). In addition, we developed a comprehensive framework of machine learning scenarios augmented with cognitive processing signals, including the practical implementation and application of these scenarios. We also discuss the ethical implications of this line of research.

Next, we transition from the goal of improving NLP models to evaluating and interpreting NLP models by leveraging cognitive signals. In particular, the prediction of cognitive signals is useful in avoiding the data scarcity limitation, while allowing us to analyze which patterns in human sentence processing are reflected in state-of-the-art language models. Moreover, we move from an anglocentric perspective to a multilingual perspective and present results covering various Indo-European languages, which increases the validity of this line of research.

Finally, we built CogniVal, the first multi-modal framework for evaluating the cognitive plausibility of pre-trained word embeddings and language models. This open source framework is

[1] https://github.com/norahollenstein/cognitiveNLP-dataCollection/wiki

available as a command line interface and allows NLP researchers and practitioners to evaluate pre-trained or custom language models against a wide range of cognitive data sources.

1.4 Impact

In the longer term, the goal of this line of research is to understand how to leverage the linguistic structure encountered in human language processing data to improve, evaluate, and interpret neural language models. The ultimate goal of this research is to advance ML methods through improving our understanding of human learning, and vice versa.

The impact of our work is twofold. First, the fundamental research conducted is beneficial for various disciplines. It will improve our understanding of ML methods as well as of the human language processing mechanisms. Understanding how humans process speech and text will provide insights into how we can improve ML algorithms to achieve similar capabilities in computational language models. The work presented within this dissertation also opens up new opportunities for correlated research hypotheses between computational neuroscience and linguistics. Second, we believe that developing generalizable and interpretable human-grounded models for language understanding will bring great advances to the field of NLP. The insights gained into the cognitive domain of language can be leveraged to build NLP applications that generalize better and are more easily interpretable and cognitively plausible. This research avenue enables more transparent machine learning, which in turn enhances the benefits of these new technologies for society and the interaction between humans and machine learning.

1.5 Organization of the Thesis

This thesis is organized as follows. In Chapter 2, we present our data collection efforts for the *Zurich Cognitive Language Processing Corpus*. In Chapter 3, we the leverage this data and other sources of eye tracking and brain activity metrics during reading to improve NLP tasks such as named entity recognition and sentiment analysis. Since brain activity data has not been explored much for these purposes, in Chapter 4, we describe a large-scale study of decoding EEG signals for machine learning in a multi-modal machine learning scenario. In Chapter 5, we move from improving NLP tasks to using human cognitive data for evaluating the

cornerstones of NLP applications, i.e., pre-trained language models. We investigate the extent to which human reading behavior can be predicted by state-of-the-art pre-trained language models. On the one hand, this alleviates the data scarcity problem when improving NLP models with cognitive signals, and on the other hand, it provides some insights into the interpretability and cognitive plausibility of language models. In Chapter 6, we present *CogniVal*, a framework for cognitive language model evaluation. Finally, in Chapter 7, we discuss the insights gained from previous chapters, their limitations, and the potential for future research directions.

1.6 Author's Publications

This dissertation is based largely on the following publications presented in chronological order:

- **Hollenstein, N.**, Rotsztejn, J., Troendle, M., Pedroni, A., Zhang, C., & Langer, N. (2018, December). ZuCo, a Simultaneous EEG and Eye Tracking Resource for Natural Sentence Reading. *Scientific data*, 5(1).
- **Hollenstein, N.**, & Zhang, C. (2019, June). Entity Recognition at First Sight: Improving NER with Eye Movement Information. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*.
- **Hollenstein, N.**, de la Torre, A., Langer, N., & Zhang, C. (2019, November). CogniVal: A Framework for Cognitive Word Embedding Evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*.
- Pfeiffer, C., **Hollenstein, N.**, Zhang, C., & Langer, N. (2020, March). Neural Dynamics of Sentiment Processing During Naturalistic Sentence Reading. *NeuroImage*, 116934.
- **Hollenstein, N.**, Troendle, M., Zhang, C., & Langer, N. (2020, May). ZuCo 2.0: A Dataset of Physiological Recordings During Natural Reading and Annotation. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*.
- **Hollenstein, N.**, Barrett, M., & Beinborn, L. (2020, May). Towards Best Practices for Leveraging Human Language Processing Signals for Natural Language Processing. In *Proceedings of the Second Workshop on Linguistic and Neurocognitive Resources (LiNCR)*.

- Barrett, M., & **Hollenstein, N.** (2020, September). Sequence Labelling and Sequence Classification with Gaze: Novel Uses of Eye-Tracking data for Natural Language Processing. *Language and Linguistics Compass*.
- **Hollenstein, N.**, van der Lek, A., & Zhang, C. (2020, December). CogniVal in Action: An Interface for Customizable Cognitive Word Embedding Evaluation. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*.
- **Hollenstein, N.**, Renggli, C., Barrett, M., Troendle, M., Langer, N., & Zhang, C. A Large-Scale Study of Decoding EEG Brain Activity for Multi-Modal Natural Language Processing. *Under review*.
- **Hollenstein, N.**, Pirovano, F., Zhang, C., Jäger, L., & Beinborn, L. Multilingual Language Models Predict Human Reading Behavior. *Under review*.

Further published work, which is outside the scope of this thesis:

- Rotsztejn, J., **Hollenstein, N.**, & Zhang, C. (2018, June). ETH-DS3Lab at SemEval-2018 Task 7: Effectively Combining Recurrent and Convolutional Neural Networks for Relation Classification and Extraction. In *Proceedings of The 12th International Workshop on Semantic Evaluation (SemEval)*.
- Barrett, M., Bingel, J., **Hollenstein, N.**, Rei, M., & Søgaard, A. (2018, October). Sequence Classification with Human Attention. In *Proceedings of the 22nd Conference on Computational Natural Language Learning (CoNLL)*.
- Girardi, I., Ji, P., Nguyen, A. P., **Hollenstein, N.**, Ivankay, A., Kuhn, L., Marchiori, C. & Zhang, C. (2018, October). Patient Risk Assessment and Warning Symptom Detection Using Deep Attention-Based Neural Networks. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*.
- Kapur, A., Sarawgi, U., Wadkins, E., Wu, M., **Hollenstein, N.**, & Maes, P. (2020, April). Non-Invasive Silent Speech Recognition in Multiple Sclerosis with Dysphonia. In *Machine Learning for Health Workshop*.
- Russo, G., **Hollenstein, N.**, Musat, C., & Zhang, C. (2020, November). Control, Generate, Augment: A Scalable Framework for Multi-Attribute Text Generation. *Findings of EMNLP*.

- Aguilar, L., Dao, D., Gan, S., Gürel, M. N., **Hollenstein, N.**, ..., Zhang, C. (2021, January). Ease.ML: A Lifecycle Management System for Machine Learning. In *Proceedings of the Conference on Innovative Database Research (CIDR)*.

Chapter 2

Collecting Cognitive Processing Signals

In order to train NLP applications, large labeled datasets are often required. For instance, to train a sentiment analysis system, which predicts the sentiment of a sentence (i.e., *positive / negative / neutral*), thousands of annotated sentences are needed. Typically, human annotators must read these training sentences and assign a sentiment to each one. Clearly, this reflects a significant investment. Our long-term goal is to replace this labor-intensive and expensive task with physiological activity data recorded from humans while reading sentences. That is to say, we aim to find and extract relevant aspects of text understanding and annotation directly from the source, i.e., eye tracking and brain activity signals during reading. By way of illustration, opinions and sentiments are elicited from a person reading text, which is reflected in their brain activity. Hence, it should be possible to decode this information from the recorded brain activity data with machine learning techniques and bypass – or at least complement – manual human annotation.

Whether it is possible to decode such information from brain activity is an empirical question and has not been answered so far. Yet, previous studies have demonstrated that eye movement information improves NLP tasks such as part-of-speech tagging (Barrett et al., 2016), sentiment analysis (Mishra et al., 2017a) and word embedding evaluation (Søgaard, 2016). In addition, there are some available resources of subjects' eye movement recordings while reading text,

The contents of this chapter are largely based on the following publications: (1) Hollenstein et al. (2018). ZuCo, a simultaneous EEG and eye tracking resource for natural sentence reading. *Scientific Data*. (2) Hollenstein et al. (2020d). ZuCo 2.0: A dataset of physiological recordings during natural reading and annotation. *LREC*.

e.g., the Dundee corpus (Kennedy et al., 2003) and the GECO corpus (Cop et al., 2017). However, while there are studies that combine electroencephalography (EEG) and eye tracking from a psycholinguistic motivation (for instance, the response to syntactically incorrect sentences (Vignali et al., 2016)), up to now there is no dataset available that combines eye tracking and brain activity that is tailored for training machine learning algorithms to perform NLP tasks.

We therefore compile a dataset that will enable researchers to advance the training of NLP applications using rich physiological data. This dataset is tailored specifically to use for cognitively-inspired NLP: We recorded as many sentences as possible for downstream machine learning applications and provide relevant features for improving and evaluating NLP models as well as for analyzing the human annotation process.

In this chapter, we first give an overview of eye tracking and EEG methods and their applications, specifically related to psycholinguistics and reading studies (Section 2.1). Then, we present the Zurich Cognitive Language Processing Corpus (ZuCo), a dataset combining electroencephalography (EEG) and eye tracking recordings from subjects reading natural sentences (Section 2.2). In total, ZuCo includes high-density EEG and eye tracking data of 30 healthy adult native English speakers, each reading natural English text for 3–6 hours. The recordings span two normal reading tasks and one task-specific reading task.

2.1 Background

Eye tracking is the process of measuring the point of gaze (fixation position of the eye) and/or eye movements. Eye trackers are used in various research disciplines including computational vision, psychology, marketing, and human-computer interaction. In psycholinguistics, eye movements are known to be indirect measurements of cognitive load during reading. Eye tracking metrics measure fixations, i.e., the maintaining of the visual gaze on a single location, and saccades, i.e., quick, simultaneous movement of both eyes between fixations.

Eye trackers provide millisecond-accurate records of where humans look when they are reading, and they are becoming cheaper and more easily available by the day (San Agustin et al., 2009; Sewell and Komogortsev, 2010). Although eye tracking data is still being recorded in controlled experiment environments, this will likely change in the near future. Recent approaches have shown substantial improvements in recording gaze data while reading by using cameras of mobile devices (Gómez-Poveda and Gaudioso, 2016; Papoutsaki et al., 2016). Hence, eye tracking

data will probably be more accessible and available in much larger volumes in due time, which will facilitate the creation of sizable datasets enormously.

Electroencephalography (EEG) refers to the recording of the brain's spontaneous electrical activity over a period of time, as recorded from multiple electrodes. It is a physiological monitoring method to record the electrical activity of the brain. EEG measures voltage fluctuations resulting from the ionic current within the neurons of the brain. It is typically non-invasive with the electrodes placed along the scalp and tells us, from the surface measurements, how active the brain is. Common clinical applications of EEG include the diagnosis of autism and epilepsy as well as sleep research. One of the biggest advantages to EEG is the ability to see brain activity as it unfolds in real time, at the level of milliseconds (high temporal resolution). However, one of the biggest disadvantages of EEG is that it is difficult to figure out where in the brain the electrical activity is coming from (low spatial resolution). EEG poorly measures neural activity that occurs below the upper layers of the brain (the cortex). Advances in EEG technology have also brought forth portable devices that allow recordings outside of the laboratory (e.g., Badcock et al. (2013)).

Some eye tracking corpora of natural reading (e.g., the Dundee corpus (Kennedy et al., 2003), Provo corpus (Luke and Christianson, 2017) and GECO corpus (Cop et al., 2017)), and a few EEG corpora (e.g., the UCL corpus by Frank and Willems (2017)) are available. It has been shown that this type of cognitive processing data is useful for improving and evaluating NLP methods (e.g., Long et al. (2017), Barrett et al. (2018a), Hale et al. (2018)). However, before the Zurich Cognitive Language Processing Corpus (ZuCo 1.0), there was no available data for simultaneous eye tracking and EEG recordings of natural reading. Dimigen et al. (2011) studied the linguistic effects of eye movements and EEG co-registration in natural reading and showed that they accurately represent lexical processing. Moreover, the simultaneous recordings are crucial to extract word-level brain activity signals.

While the above-mentioned studies analyze and leverage natural reading, some NLP work has used eye tracking during annotation (but, as of yet, not EEG data). Mishra et al. (2016a) and Joshi et al. (2014) recorded eye tracking during binary sentiment annotation (positive/negative). This data was used to determine the annotation complexity of the text passages based on eye movement metrics (Mishra et al., 2017b) and for sarcasm detection (Mishra et al., 2016b). Moreover, eye tracking has been used to analyze the word sense annotation process in Hindi (Joshi et al., 2013), named entity annotation in Japanese (Tokunaga et al., 2017), and to leverage annotator gaze behaviour for English co-reference resolution (Cheri et al., 2016). Finally, Tomanek et al. (2010) used eye tracking data during entity annotation to build a cost model for active

learning. However, until now there was no available data or research that analyzes the differences in the human processing of normal reading versus reading during information searching, i.e., annotation.

Importance of Naturalistic Recordings

A prominent feature of the datasets we present in this chapter is the personal reading speed. The sentences were presented to the subjects in a naturalistic reading scenario, where the complete sentence is presented on the screen and the subjects read each sentence at their own speed, i.e., the reader determines themselves for how long each word is fixated and which word to fixate next. Additionally, a naturalistic experiment setup also involves the selection of the stimuli. While most previous reading studies use carefully constructed stimuli to analyze specific linguistic phenomena, recent methodological advances allow for recordings of naturalistic language and measuring how the brain responds to language as it is used (Alday, 2019; Hamilton and Huth, 2018). With the purpose of recording the reading process as unconstrained as possible, we selected stimulus sentences that naturally occur on the web.

Even though eye tracking and EEG signals are still being recorded in controlled experiment environments, this will likely change in the near future. Recent approaches have shown great improvements in recording eye tracking data using the web-cams of mobile devices (Gómez-Poveda and Gaudioso, 2016; Papoutsaki et al., 2016). Moreover, collecting EEG data in everyday situations is also becoming more feasible with portable commercial devices (e.g., Stytsenko et al. (2011). Mostow et al. (2011), for instance, propose using simple single-channel EEG headsets during class in school to analyze reading comprehension. Nevertheless, there is still a wide gap between recording EEG data in natural environments and using it successfully in machine learning applications, since the EEG signals have a low signal-to-noise ratio and it is yet unclear in which features the signal is concentrated (see Chapter 4). Moreover, EEG quality varies greatly from subject to subject, resulting in a large variance in the performance of machine learning models trained on these data.

Importance of Co-Registration Studies

For some neurolinguistic corpora, data from co-registration studies is available, which means two modalities were recorded simultaneously during the same experiment (Degno and Liv-

ersedge, 2020). This has become more popular, since the recording modalities are complementary in terms of temporal and spatial resolution as well as the directness in the measurement of neural activity (Mulert, 2013). Brain-electric correlates of reading have traditionally been studied with word-by-word presentation, a condition that eliminates important aspects of the normal reading process and precludes direct comparisons between neural activity and oculomotor behavior. Dimigen et al. (2011) presented a study of simultaneous EEG and eye movement recordings of subjects reading German sentences. This work focuses on demonstrating the benefits of simultaneous EEG and eye movement recording. The authors show that EEG indices of semantic processing can be obtained in natural reading and compared to eye movement behavior.

Recent reports attest to the feasibility of co-registration studies for studying the neurobiology of natural reading (see Kandylaki and Bornkessel-Schlesewsky (2019) for a review). For example, eye tracking and EEG recorded concurrently during reading (Dimigen et al., 2011; Henderson et al., 2013) and concurrent eye tracking and fMRI (Henderson et al., 2015, 2016). Thus, it is important to emphasize the value of the simultaneous EEG and eye tracking recordings of the datasets presented in this work (Kandylaki and Bornkessel-Schlesewsky, 2019). The high temporal resolution of EEG signals in combination with eye tracking signals permits us to define exact word boundaries in the timeline of a subject reading a sentence, allowing to extract brain activity signals for each word. For machine learning applications, using data from co-registration studies in NLP allows for comparisons on the same language stimuli, on the same population, and on the same language understanding task, where only the recording method differs.

2.2 Zurich Cognitive Language Processing Corpus

In this section, we describe the compilation of the Zurich Cognitive Language Processing Corpus (ZuCo). ZuCo is a dataset combining electroencephalography (EEG) and eye tracking recordings from subjects reading natural sentences. ZuCo includes high-density EEG and eye tracking data of 30 healthy adult native English speakers, each reading natural English text for 3–6 hours. The recordings span two normal reading tasks and one task-specific reading task. We recorded two datasets. The first part, ZuCo 1.0, encompasses EEG and eye tracking data of 21,629 words in 1,107 sentences. The second part, ZuCo 2.0, encompasses the same type of recordings of 15,138 words and 739 sentences.

The main difference and reason for recording ZuCo 2.0 consists in the experiment procedure,

namely, the number of sessions and the order of the reading tasks. For ZuCo 1.0, the normal reading and task-specific reading paradigms were recorded in different sessions on different days. Therefore, the recorded data is not fully appropriate as a means of comparison between natural reading and annotation, since the differences in the brain activity data might result mostly from the different sessions due to the sensitivity of EEG. This, and extending the dataset with more sentences and more subjects, were the main factors for recording the current corpus. We purposefully maintained an overlap of some sentences between both datasets to allow additional analyses (details are described in Section 2.2.1). If not stated explicitly, the procedures described below apply to both parts of the ZuCo corpus. Both datasets, including the raw data and the extracted features, are freely available on the Open Science Framework.[1]

2.2.1 Experimental Setup

In this section, we describe the experimental setup designed for recording the Zurich Cognitive Language Processing Corpus, including the selection of participants and their linguistic assessments, as well as the design of the reading paradigms and the data acquisition details. Please note that, if not distinctly stated, all descriptions apply to ZuCo 1.0 and ZuCo 2.0 equally.

Participants

For ZuCo 1.0, data were recorded from 12 healthy adults, all native English speakers (originating from Canada, USA, UK, or Australia). The study included 5 female and 7 male subjects, all right-handed, of age between 22 and 54 years. Details about the participants' age and gender can be found in Table 2.1. All participants gave written consent for their participation and the re-use of the data prior to the start of the experiments. The study was approved by the Ethics Commission of the University of Zurich.

For ZuCo 2.0, we recorded data from 19 participants and discarded the data of one of them due to technical difficulties with the eye tracking calibration. Hence, we share the data of 18 participants. All participants are healthy adults (between 23 and 52 years old; 10 females). Their native language is English, originating from Australia, Canada, UK, USA or South Africa.

[1] ZuCo 1.0: `https://osf.io/q3zws/` and ZuCo 2.0: `https://osf.io/2urht/`.

ID	Age	Gender	LexTALE	Comprehension Scores			Reading Speed		
				SR	NR	TSR	SR	NR	TSR
ZAB	41	female	100.00%	76.09%	86.11%	90.42%	4.88	5.14	3.32
ZDM	25	male	100.00%	76.09%	80.56%	96.81%	4.41	5.13	3.32
ZDN	32	male	97.50%	89.13%	86.11%	92.87%	3.91	4.10	2.93
ZGW	49	male	91.25%	71.74%	86.11%	92.14%	6.87	8.06	4.17
ZJM	41	male	77.50%	80.43%	97.22%	96.56%	6.22	8.73	6.30
ZJN	51	female	97.50%	54.34%	83.33%	79.12%	8.71	11.30	7.10
ZJS	42	male	97.50%	91.30%	91.67%	93.86%	4.34	4.18	2.88
ZKB	26	female	100.00%	89.13%	86.11%	95.33%	5.39	8.43	2.48
ZKH	41	female	81.25%	76.09%	83.33%	93.12%	5.42	6.43	5.57
ZKW	25	female	96.25%	69.57%	91.67%	94.84%	6.94	11.73	6.14
ZMG	51	male	100.00%	91.30%	88.89%	95.82%	4.39	5.33	3.73
ZPH	26	male	97.50%	89.13%	94.44%	97.05%	4.78	7.55	2.71
mean	**38**	**-**	**94.69%**	**79.53%**	**87.96%**	**93.16%**	**5.52**	**7.18**	**4.22**

Table 2.1: Subject demographics for ZuCo 1.0, LexTALE scores, scores of the comprehension questions, and individual reading speed (i.e., seconds per sentence) for each task.

Two participants are left-handed and three participants wear glasses for reading. Details on subject demographics can be found in Table 2.2. All participants gave written consent for their participation and the re-use of the data prior to the start of the experiments. The study was conducted under the same approval by the Ethics Commission of the University of Zurich as ZuCo 1.0.

Linguistic Assessment

The vocabulary and language proficiency of all participants was tested with the LexTALE test (Lexical Test for Advanced Learners of English; Lemhöfer and Broersma (2012)). LexTALE is an unspeeded lexical decision task, which is for intermediate to highly proficient language users. Table 2.1 shows the detailed scores per subject for ZuCo 1.0, and Table 2.2 the scores for the participants of ZuCo 2.0. These scores are the percentages of correctly answered control questions in the respective task. The average score of all subjects in ZuCo 1.0 on the LexTALE test was 94.69% and of all subjects in ZuCo 2.0 it was 88.54%.

ID	Age	Gender	LexTALE	Comprehension Scores		Reading Speed	
				NR	TSR	NR	TSR
YAC	32	female	76.25%	82.61%	83.85%	5.27	4.96
YAG	47	female	93.75%	91.30%	56.92%	7.64	8.73
YAK	31	female	100.00%	74.07%	96.41%	3.83	5.89
YDG	51	male	100.00%	91.30%	96.67%	4.97	3.93
YDR	25	male	85.00%	78.26%	96.92%	4.32	2.32
YFR	27	male	85.00%	89.13%	94.36%	6.48	4.79
YFS	39	male	90.00%	91.30%	96.15%	3.96	2.85
YHS	31	male	90.00%	78.26%	97.69%	3.30	2.40
YIS	52	male	97.50%	89.13%	98.46%	5.82	2.58
YLS	34	female	93.75%	91.30%	92.31%	5.57	5.85
YMD	31	female	100.00%	86.96%	95.64%	7.50	6.24
YMS	36	female	86.25%	89.13%	95.38%	7.68	3.35
YRH	28	female	81.25%	86.96%	95.64%	5.14	4.32
YRK	29	female	85.00%	97.83%	96.15%	7.35	7.70
YRP	23	female	82.50%	78.26%	90.00%	7.14	8.37
YSD	34	male	95.00%	93.48%	94.36%	5.01	2.87
YSL	32	female	71.25%	84.78%	83.85%	6.73	6.14
YTL*	36	male	81.25%	80.43%	94.10%	7.48	3.23
mean	**34**	**44% m.**	**88.54%**	**86.36%**	**91.94%**	**5.84**	**4.81**

Table 2.2: Subject demographics for ZuCo 2.0, LexTALE scores, scores of the comprehension questions, and individual reading speed (i.e.,seconds per sentence) for each task. The * next to the subject ID marks a bilingual subject.

Reading Materials

The dataset we present is composed of sentences from the Stanford Sentiment Treebank (Socher et al., 2013) and the Wikipedia relation extraction corpus (Culotta et al., 2006). The Stanford Sentiment Treebank contains single sentences extracted from movie reviews with manually annotated sentiment labels. For ZuCo 1.0, we randomly selected 400 very positive, very negative, or neutral sentences (4% of the full treebank). The 400 selected sentences are comprised of 123 neutral, 137 negative, and 140 positive sentences. ZuCo 2.0 does not contain any sentiment

sentences. The Wikipedia relation extraction dataset contains paragraphs about famous people, which were labeled with relation types.

For ZuCo 1.0, we extracted the following data subsets: For the normal reading we randomly selected 650 sentences that contain a relation (14% of the full dataset). For the task-specific relation extraction reading we selected approximately 40 sentences of each of the following relation types: *award*, *education*, *employer*, *founder*, *job_title*, *nationality*, *political_affiliation*, *visited* and *wife*. Of these sentences, 300 were used in the normal reading tasks and 407 in the task-specific task. Note that 48 sentences are duplicates and appear in both tasks. These duplicate sentences can be found in a separate file in the online repository. The dataset statistics are shown in Table 2.3.

For ZuCo 2.0, the participants read 739 sentences that were selected from the Wikipedia corpus provided by Culotta et al. (2006) during a single recording session. This corpus was chosen because it provides annotations of semantic relations. Relation detection is a high-level semantic task requiring complex cognitive processing. We included seven of the originally defined relation types: *political_affiliation*, *education*, *founder*, *wife/husband*, *job_title*, *nationality*, and *employer*. The sentences were chosen in the same length range as ZuCo 1.0, and with similar Flesch reading ease scores (Kincaid et al., 1975). The dataset statistics are shown in Table 2.4. Of the 739 sentences, the participants read 349 sentences in a normal reading paradigm and 390 sentences in a task-specific reading paradigm, in which they had to determine whether a certain relation type occurred in the sentence or not. Table 2.5 shows the distribution of the different relation types in the sentences of the task-specific annotation paradigm of both ZuCo 1.0 and ZuCo 2.0.

In ZuCo 2.0, there purposefully are 63 duplicates between the normal reading and the task-specific sentences (8% of all sentences). The intention of these duplicate sentences is to provide a set of sentences read twice by all participants with a different task in mind. This enables the comparison of eye tracking and brain activity data when reading normally and when annotating specific relations (see examples in Appendix A.2).

Furthermore, there is also an overlap in the sentences between ZuCo 1.0 and ZuCo 2.0. 100 normal reading and 85 task-specific sentences recorded for this dataset were already recorded in ZuCo 1.0. This allows for comparisons between the different recording procedures (i.e.,session-specific effects) and between more participants (i.e., subject-specific effects).

Table 2.6 presents the overall descriptive statistics of the materials split by task. Sentences from both parts of the dataset were presented to the subjects in three different tasks of normal reading

	Normal Reading (Sentiment)	**Normal Reading** (Wikipedia)	**Task-specific Reading** (Wikipedia)	**Total**
Words	7,079	6,386	8,164	21,629
Word types	3,080	2,657	2,995	7,099
Sentences	400	300	407	1,107
Flesch score	56.71	51.33	51.43	53.16
Sent. length	17.7 (8.3), 3-43	21.3 (10.6), 5-62	20.1 (10.1), 5-62	19.5 (9.7), 3-62
Word length	7.0 (2.7), 1-26	6.7 (2.7), 1-29	6.7 (2.6), 1-21	6.8 (2.7), 1-29

Table 2.3: Descriptive statistics of the reading materials of ZuCo 1.0. We report mean (standard deviation), and range for sentence length and word length.

	Normal Reading (Wikipedia)	**Task-specific Reading** (Wikipedia)	**Total**
Words	6,828	8,310	15,138
Word types	2,412	2,437	3,789
Sentences	349	390	739
Flesch score	55.38	50.76	53.07
Sent. length	19.6 (8.8), 5-53	21.3 (9.5), 5-53	20.6 (9.4), 5-53
Word length	4.9 (2.7), 1-29	4.9 (2.7), 1-29	4.9 (2.7),1-29

Table 2.4: Descriptive statistics of the reading materials of ZuCo 2.0. We report mean (standard deviation), and range for sentence length and word length.

and task-specific reading. All sentences can be also found in the MATLAB files with the eye tracking and EEG data, and in separate text files in the online repository. These additionally include the original sentiment and relation labels.

Experiment Design

During all tasks of ZuCo 1.0 and 2.0, the sentences were presented one at a time at the same position on the screen. Text was presented in black with font size 20-point Arial on a light grey background resulting in a letter height of 0.8 mm or 0.674°. The lines were triple-spaced, and the words double-spaced. A maximum of 80 letters or 13 words were presented per line in all three

Relation type	ZuCo 1.0	ZuCo 2.0
Political affiliation	38 (9)	45 (9)
Education	46 (9)	72 (10)
Wife/Husband	55 (8)	54 (12)
Job title	40 (9)	65 (11)
Employer	40 (10)	54 (10)
Nationality	40 (9)	60 (8)
Founder	42 (8)	40 (8)
Award	60 (8)	0
Visited	41 (11)	0
Total	**407 (81)**	**390 (68)**

Table 2.5: Distribution of relation types in the task-specific reading task in ZuCo 2.0. The right column contains the number of sentences, and the number control sentences without a relation in brackets.

tasks. Long sentences spanned multiple lines. A maximum of 7 lines for the Sentiment normal reading task, 5 lines for the Wikipedia normal reading tasks, and 7 lines for the task-specific reading paradigm were presented simultaneously on the screen.

In all tasks of both ZuCo 1.0 and ZuCo 2.0, the subjects used a control pad to switch to the next sentence and to answer the control questions, which allowed for natural reading. Compared to RSVP (Rapid Serial Visual Representation), where each word is presented separately at an equal speed, the normal reading approach is closer to a natural reading scenario: The subjects read each sentence at their own speed, i.e., the reader determines for how long each word is fixated and which word to fixate next. This allows for varying reading speed between the subjects; each subject reads at his/her own personal pace. Tables 2.1 and 2.2 show the average reading speed, i.e., the average number of seconds a subject spends per sentence, reported for each task. Note that a Wilcoxon test revealed a significant difference in reading speed between the NR and TSR tasks ($Z = 3.0594$; $p < 0.01$; see Figure 2.1 for the distribution of the reading speed per task for both ZuCo 1.0 and ZuCo 2.0). The individual reading speed for every subject was considerably lower during the TSR task than during the NR task. Although the reading material was of the same type, in the NR task passive reading was recorded, while during the TSR task the subjects had to search for a specific relation type. Thus, the task-specific reading led to shorter passes because the goal was merely to recognize a relation in the text, but not necessarily to process the whole meaning of the sentences. The task instructions are described in detail below.

	Normal reading (SR) (Sentiment)	**Normal reading (NR)** (Wikipedia)	**Task-specific reading (TSR)** (Wikipedia)
Material	Positive, negative or neutral sentences from movie reviews	Wikipedia sentences containing specific relations	Wikipedia sentences containing specific relations
Example	"The film often achieves a mesmerizing poetry." (positive)	"Talia Shire (born April 25, 1946) is an American actress of Italian descent." (relations: nationality, job title)	"Lincoln was the first Republican president." (relation: political affiliation)
Task	Read the sentences, rate the quality of the movie based on the sentence read	Read the sentences, answer control questions	Mark whether a specific relation occurs in the given sentence or not
Control	"Based on the previous sentence, how would you rate this movie from 1 (very bad) to 5 (very good)?"	"Talia Shire was a ... 1) singer 2) actress 3) director"	"Does this sentence contain the political affiliation relation? 1) Yes 2) No"

Table 2.6: Schematic overview of the three tasks in the ZuCo experiment design. ZuCo 1.0 contains all three tasks, ZuCo 2.0. contains only NR and TSR.

Sentiment Normal Reading (SR) In the first task, the subjects were presented with positive, negative, or neutral sentences for normal reading to analyze the elicitation of emotions and opinions during reading. Figure 2.2 (left) shows a sample sentence and how it was presented on the screen. The movie ratings in the control condition questions were answered with the numbers 1-5 (very bad - very good) on the control pad. The task was explained to the subject orally, followed by instructions on the screen.

As a control condition, the subjects had to rate the quality of the described movies in 47 of the 400 sentences. The average response accuracy compared to the original labels of the Stanford Sentiment Treebank is 79.53%, and the response accuracy per subject can be found in Table 2.1.

Wikipedia Normal Reading (NR) In the second task, the subjects were presented with sentences that contained semantic relations. The sentences were presented to the subject in the same manner as in the previous task. The numbers on the control pad were used to choose the answers of the control questions. Figure 2.2 (middle) shows an exemplary sentence as it appeared on the

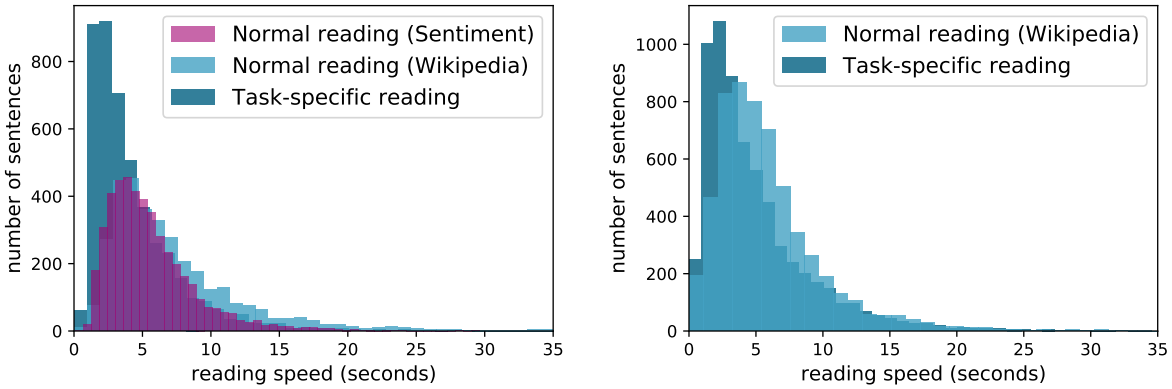


Figure 2.1: Histogram of the reading speed of all sentences read by all subjects in ZuCo 1.0 (left) and ZuCo 2.0 (right).

screen. The task was explained to the subjects orally, followed by instructions on the screen.

The control condition for this task consisted of multiple-choice questions about the content of the previous sentence. 12% of randomly selected sentences were followed by such a comprehension question with three answer options on a separate screen.The average response accuracy in ZuCo 1.0 is 87.96%, and the response accuracy per subject can be found in Table 2.3. The average response accuracy in ZuCo 2.0 is 86.36% and the response accuracy per subject can be found in Table 2.4.

Task-specific Reading (TSR) In a subsequent session, the subjects were presented with similar sentences as in the second task, but with specific instructions to focus on a certain relation type. This allows us to compare the EEG and eye tracking signals during normal reading to reading with a specific relation in mind. As a control condition, the subjects had to report for each sentence whether a specific relation was present in the sentence or not. 17% of the sentences did not include the relation type and were used as control conditions.

The sentences were presented in blocks of the same relations, so the subjects would know which relation to look for without having to read the questions. Each of these blocks had a separate practice round with a definition of the relations and three sample sentences. Figure 2.2 (right) shows a sample sentence on a screen for this specific task. Note that the control question in the bottom right was presented for each sentence. The task was explained to the subjects orally, followed by instructions on the screen, including a definition of the relation type relevant in the current block of sentences.

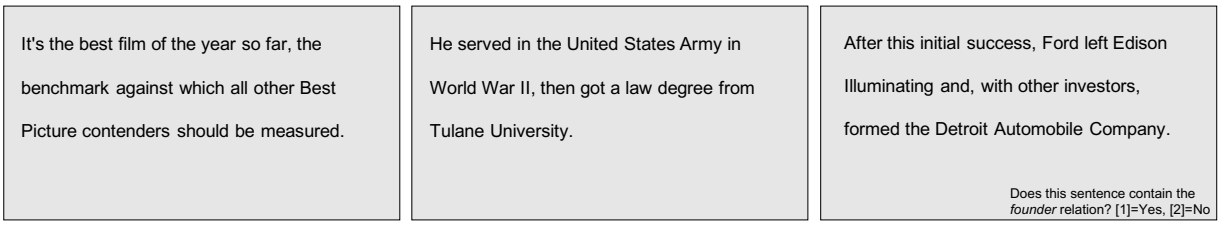


Figure 2.2: Example sentences on the recording screen for the three reading tasks included in ZuCo: (left) a normal reading sentence from the Stanford Sentiment Treebank, (middle) a normal reading sentence from the Wikipedia dataset, and (right) a task-specific reading sentence from the Wikipedia dataset.

For ZuCo 1.0, the average response accuracy on this control condition is 93.16% and the response accuracy per subject can be found in Table 2.1. For ZuCo 2.0, the average response accuracy on this control condition is 91.94% and the response accuracy per subject can be found in Table 2.2.

Recording Procedure

For ZuCo 1.0, each participant read the entire reading material in two sessions of 2–3 hours each (at the same time of day). In the first session, the participants completed Task NR, followed by the first half of Task SR. In the second session, the participants completed Task TSR, followed by the second half of Task SR. The sentences were presented to all subjects in identical order. Before each task, a practice round of 3–5 sentences was displayed for the subject to get familiar with the task. The eye tracking device was re-calibrated in blocks of 60 sentences (approximately every 10–15 minutes) for the first two tasks of normal reading and after every 40 sentences in the third task.

For ZuCo 2.0, all 739 sentences were recorded in a single session for each participant. The duration of the recording sessions was between 100 and 180 minutes, depending on the time required to set up and calibrate the devices, and the personal reading speed of the participants. We recorded 14 blocks of approx. 50 sentences, alternating between tasks: 50 sentences of normal reading, followed by 50 sentences of task-specific reading. The order of blocks and sentences within blocks was identical for all subjects. Each sentence block was preceded by a practice round of three sentences and followed by a short break to ensure a clear separation between the reading tasks.

Data Acquisition

The data acquisition was identical for both ZuCo 1.0 and ZuCo 2.0. Data acquisition occurred in the same sound-attenuated and dark experiment Faraday recording cage. Participants were seated at a distance of 68 cm from a 24-inch monitor (ASUS ROG, Swift PG248Q, display dimensions 531.4x298.9mm, resolution 800x600 pixels (resulting in a display: 400x298.9mm), vertical refresh rate of 100 Hz). The sentences were presented in black on a light grey background with font size 20-point Arial, resulting in a letter height of 0.8mm. A stable head position was ensured via a chin rest. Subjects were instructed to stay as still as possible during the tasks to avoid motor EEG artifacts. Participants were also offered snacks and water during the breaks and were encouraged to rest.

The experiment was programmed in MATLAB 2016b, using the PsychToolbox extension (Pelli, 1997; Brainard, 1997). The order of the reading paradigms was the same for all participants. Instructions for the tasks were presented on the computer screen. Participants completed the tasks sitting alone in the room, while two research assistants were monitoring their progress in the adjoining room. All recording scripts including the detailed participant instructions are available alongside the data.

Eye tracking acquisition During all of the reading paradigms, eye position and pupil size were recorded with an infrared video-based eye tracker (EyeLink 1000 Plus, SR Research, `http://www.sr-research.com/`) at a sampling rate of 500 Hz and an instrumental spatial resolution of $< 0.01°$. The eye tracker was calibrated with a 9-point grid before each paradigm. Specifically, participants were asked to direct their gaze in turn to a dot presented at each of nine locations in a random order. In a validation step, the calibration was repeated until the error between two measurements at any point was less than 0.5°, or the average error for all points was less than 1°. The calibration was re-validated before each block of sentences. When necessary, a re-calibration would also be conducted.

EEG acquisition High-density EEG data were recorded during all tasks at a sampling rate of 500 Hz with a bandpass of 0.1 to 100 Hz, using a 128-channel EEG Geodesic Hydrocel system (Electrical Geodesics, Eugene, Oregon). The recording reference was at Cz. For each participant, head circumference was measured, and an appropriately sized EEG net was selected. The impedance of each electrode was checked prior to recording to ensure good contact, and was

kept below 40 kOhm. Electrode impedance levels were checked after every third block of 60 sentences (approximately every 30 minutes) and reduced if necessary.

2.2.2 Preprocessing & Feature Extraction

Eye Tracking

The EyeLink 1000 tracker processes eye position data, identifying saccades, fixations and blinks. Saccades are detected by the velocity and acceleration of the eye movements. Here, SR-research default system parameters have been used to define saccades: an acceleration threshold of 8000° per sec^2, a velocity threshold of 30° per sec, and a deflection threshold of 0.1°. Fixations were defined as time periods without saccades. The dataset therefore consists of (x, y) gaze location entries for individual fixations (see Figure 2.3 (top)). Coordinates were given in pixels with respect to the monitor coordinates (the upper left corner of the screen was (0,0) and bottom/right was positive). We also provide raw sample data that can be used to validate the fixation detection settings. Furthermore, a blink can be regarded as a special case of a fixation, where the pupil diameter is either zero or outside a dynamically computed valid pupil, or the horizontal and vertical gaze positions are zero.

For later analysis, only fixations within the boundaries of each displayed word have been extracted. On the x-axis, the word boundaries were extended so that they were adjacent (Figure 2.3). A Gaussian mixture model was trained on the (y-axis) gaze data for each sentence to improve the allocation of the fixations to the corresponding text lines. The number of text lines determined the number of Gaussians to be fitted within the model. Subsequently, each gaze data point was clustered to the matching Gaussian and the data were realigned. As a result, each gaze data point is clearly assigned to a specific text line. Data points distinctly not associated with reading (minimum distance of 50 pixels to the text) were excluded.

On the basis of a previous eye tracking corpus by Cop et al. (2017) we have extracted the following eye tracking features in MATLAB: (1) gaze duration (GD), the sum of all fixations on the current word in the first-pass reading before the eye moves out of the word; (2) total reading time (TRT), the sum of all fixation durations on the current word, including regressions; (3) first fixation duration (FFD), the duration of the first fixation on the prevailing word; (4) single fixation duration (SFD), the duration of the first and only fixation on the current word; and (5) go-past time (GPT), the sum of all fixations prior to progressing to the right of the current

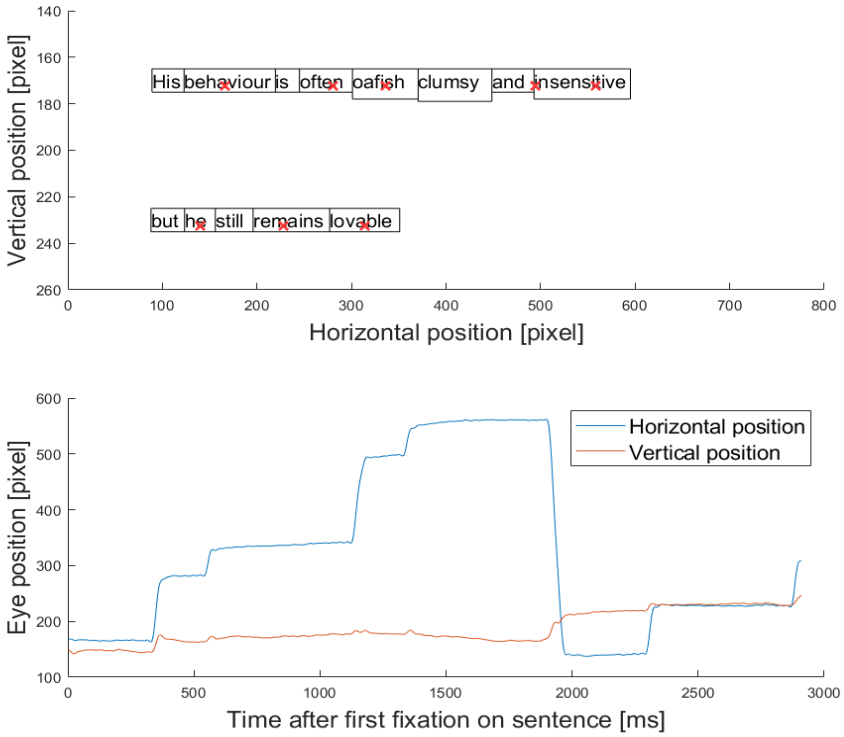


Figure 2.3: Visualization of single trial eye tracking data. (top) Prototypical single sentence fixation data for a representative subject. Red crosses indicate fixations. Boxes around the words indicate the area in which fixations are allocated to the specific word. (bottom) Raw gaze data of the fixation data plotted above.

word, including regressions to previous words that originated from the current word. For each of these eye tracking features, we have additionally computed the pupil size. Furthermore, we have extracted the number of fixations and mean pupil size for each word and sentence. Fixations that were shorter than 100 ms were excluded from the analyses, because these are unlikely to reflect fixations relevant for sentence processing (Sereno and Rayner, 2003). However, the raw eye tracking data are available to assess further potential eye tracking features.

Electroencephalography (EEG)

The data shared in this project are available as raw data, but also preprocessed with Automagic (Pedroni et al. (2019); version: 1.4.6).[2] The data from each paradigm are saved as a separate file

[2]The MATLAB code for the preprocessing can be found at `https://github.com/methlabUZH/automagic`.

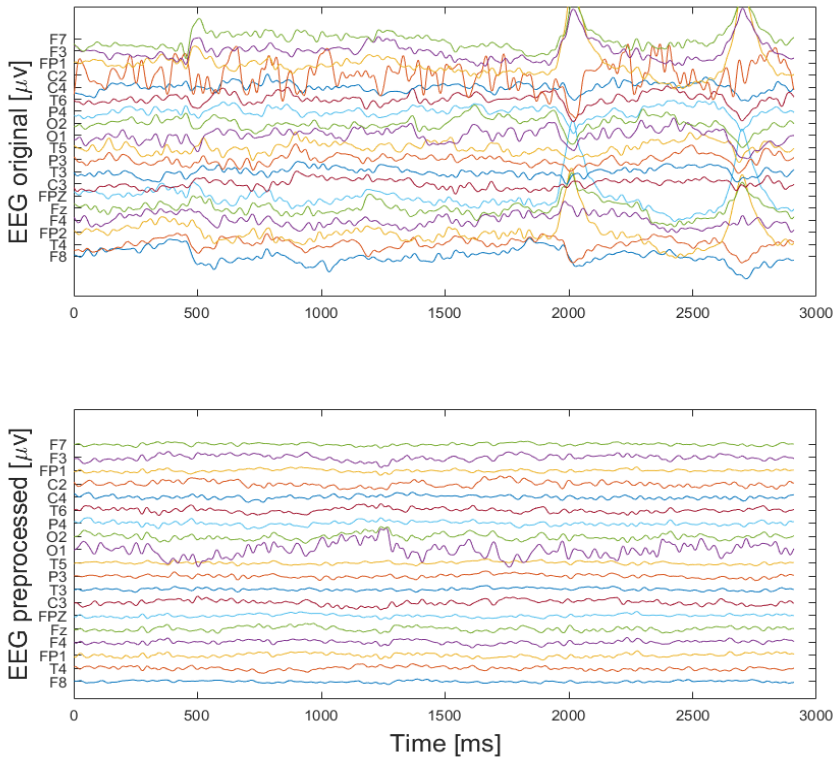


Figure 2.4: Visualization of single trial EEG data. (top) Subset of the raw EEG data during the sentence. Electrodes matching the 10–20 systems were chosen and for plotting purposes data were bandpass-filtered (0.5–30 Hz.). (bottom) Same data as in (top) after preprocessing. The y-axis labels represent specific electrode channel names.

for each participant. In the first step of preprocessing, EEG data were imported in MATLAB (`pop_readegi.m`) and the triggers and latencies for each paradigm were extracted. One hundred and five EEG channels were used for scalp recordings and nine Electro-oculography (EOG) channels were used for artifact removal. The rest of the channels lying mainly on the neck and face were discarded before data analysis.

Electrodes yielding bad quality signals were identified and replaced. Identification of bad electrodes was based on the EEGLab plugin `clean_rawdata`.[3] This plugin removes flat-line channels, low-frequency and noisy channels. A channel was defined as a bad electrode when recorded data from that electrode were correlated at less than 0.85 to an estimate based on other channels (channel criterion). Furthermore, a channel was defined as a bad channel if it had more line

[3] `http://sccn.ucsd.edu/wiki/Plugin_list_process`

noise relative to its signal compared to all other channels (4 standard deviations). Finally, if a channel had a longer flat-line than 5 s, it was considered as bad.

In a next step, the EEG data were high-pass filtered at 0.5 and notch filtered (49–51 Hz) with a Hamming windowed-sync finite impulse response zero-phase filter.[4] The filter order was defined to be 25% of the lower passband edge. Eye artifacts were removed by linearly regressing the Electrooculogpraphy (EOG) channels from the scalp EEG channels (Parra et al., 2005). The EOG electrodes were placed on the participant's forehead, outer and inner canthi (electrode numbers 8, 14, 17, 21, 25, 125, 126, 127, and 128 from the HydroCel Geodesic Sensor Net). In this study, MARA (Multiple Artifact Rejection Algorithm, Winkler et al. (2011)), a supervised machine learning algorithm that evaluates ICA components, is used for automatic artifact rejection. MARA has been trained on manual component classifications, and so captures the wide range of artifacts that manual rejection detects. MARA has proven especially effective at detecting and removing eye and muscle artifact components. Specifically, MARA evaluates each component on the six algorithm features: Current Density Norm and Range Within Pattern, Fit Error k, 8–13 Hz, and Mean Local Skewness as the feature set. MARA rejects any components with artifact probabilities greater than 0.52. Subsequently, bad electrodes were interpolated by using a spherical spline interpolation `eeg_interp.m`. Moreover, after automatic scanning, noisy channels were selected by visual inspection and interpolated. The effect of preprocessing is displayed in Figure 2.4 for a representative subject and sentence.

After preprocessing, the EEG and eye tracking data were synchronized using the "EYE EEG extension" by Winkler et al. (2014) to enable EEG analyses time-locked to the onset of the fixations. The synchronization is performed by identifying "shared" events and fitting a linear function to the shared event latencies in order to refine the estimation of the latency of the start and end events. Synchronization quality was ensured by comparing the trigger latencies recorded in the EEG and eye tracker data. All synchronization errors did not exceed one sample (2 ms).

For the purposes of the current work, we were interested in oscillatory power in different frequency bands; however, the time-series data are shared as well. To compute oscillatory power measures, we band-pass filtered the continuous EEG signals across an entire task period for five different frequency bands resulting in a time-series for each frequency band. The independent frequency bands were determined following: theta$_1$ (4–6 Hz), theta$_2$ (6.5–8 Hz) alpha$_1$ (8.5–10

[4] EEGLAB function `pop_eegfiltnew.m`

Hz), alpha$_2$ (10.5–13 Hz), beta$_1$ (13.5–18 Hz) beta$_2$ (18.5–30 Hz) and gamma$_1$ (30.5–40 Hz) and gamma$_2$ (40–49.5 Hz). We then applied a Hilbert transform to each of these time series, resulting in a complex time series. The Hilbert phase and amplitude estimation method yields results equivalent to sliding window Fast Fourier transform and wavelet approaches (Bruns, 2004). We specifically chose the Hilbert transformation to maintain temporal information for the amplitude of the frequency bands to enable the power of the different frequencies for time segments defined through fixations from the eye tracking recording. Thus, for each eye tracking feature, we computed the corresponding EEG feature in each frequency band. Furthermore, we have extracted EEG features based on sentence-level by calculating the power in each frequency band. For all EEG features, we have additionally calculated the difference of the power spectra between the frontal left and right homologue electrodes pairs. For each EEG eye tracking feature, all channels were subject to an artifact rejection criterion of $\pm 90 \mu V$ to exclude trials with transient noise.

2.3 Discussion

The technical validation that proves the quality of the recorded data can be found in Appendix A.2. We show that the reading time features as well as the fixation-related potentials in the EEG data are comparable to the results in previous studies and also that the recordings of ZuCo 1.0 and ZuCo 2.0 provide the same feature ranges and recording quality. We further show the differences between normal reading and task-specific annotation reading, which is apparent in the eye tracking data, but also evident in the brain activity recordings (see Figure 2.5).

We want to highlight the re-use potential of this data. It allows to conduct experiments for different NLP tasks. Possible NLP applications are information extraction for text mining, including entity and relation discovery, and semantic tasks, such as sentiment analysis. To train machine learning systems, the number of samples (i.e.,words and sentences) is crucial. Hence, in this work we focused more on the number of sentences recorded than the number of subjects. While this dataset has been created with machine learning and natural language processing as its primary application, this data can also be used to analyze the human reading process from a neuroscience perspective. It can be used for linguistic and (neuro-)psychological studies to generate new hypotheses (exploratory analyses), but these hypotheses should then be tested on a higher number of subjects to account for the variability of reading strategies across subjects.

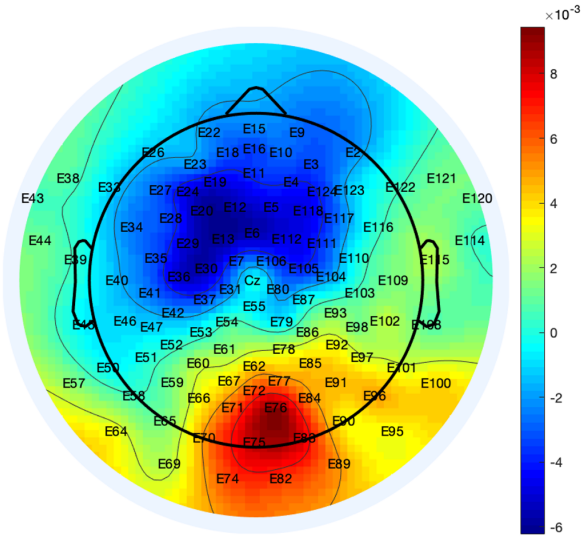


Figure 2.5: Topography plot of the voltage differences in the EEG activity between both reading tasks, averaged over all sentences all all subjects of the dataset. The figure shows the scalp viewed from above, nose at the top, with all electrode names at the position of recording.

2.4 Summary

We presented a new, freely available corpus of eye movements and electrical brain activity recordings during natural reading and during annotation. This is the first dataset that allows for the comparison between these two reading paradigms. We described the materials and experiment design in detail and conducted an extensive validation to ensure the quality of the recorded data. Since this corpus is tailored to cognitively-inspired NLP, the applications and re-use potentials of this data are extensive. The provided word-level *and* sentence-level eye tracking and EEG features can be used to improve and evaluate NLP and machine learning methods, for instance, to evaluate linguistic phenomena in neural models via neurolinguistic data. For instance, human language processing recordings can be used to probe the syntactic skills of language models (Toneva and Wehbe, 2019) or to evaluate the cognitive plausibility of word representations (see Chapter 6). In addition, because the sentences contain semantic relation labels and the annotations of all participants, it can also be widely used for relation extraction and classification. Finally, the two carefully constructed reading paradigms allow for the comparison between normal reading and reading during annotation, which can be relevant to improve the manual labelling process as well as the quality of the annotations for supervised ML.

We believe that this dataset represents a valuable resource for natural language processing. As we will show in Chapter 3, the EEG and eye tracking signals lend themselves to train improved machine-learning models for various tasks, in particular for information extraction tasks such as entity and relation extraction and sentiment analysis. Moreover, this dataset is useful for advancing research into the human reading and language understanding process at the level of brain activity and eye movement.

Chapter 3

Improving Natural Language Understanding with Cognitive Signals

In this chapter, we leverage the ZuCo data described in Chapter 2, as well as other available cognitive data sources, to improve machine learning models for natural language understanding (NLU). First, we present the related work in this area of research in Section 3.1. In Section 3.2, we conducted experiments leveraging eye tracking data to improve a neural models of named entity recognition (NER) as a first exploration of this topic. Previous research shows that eye tracking data contains information about the lexical and syntactic properties of text, which can be used to improve natural language processing models. We leverage eye movement features from three corpora with recorded gaze information to augment a state-of-the-art neural model for NER with gaze embeddings. These corpora were manually annotated with named entity labels. Moreover, we show how gaze features, generalized on word type level, eliminate the need for recorded eye tracking data at test time. The gaze-augmented models for NER using token-level and type-level features outperform the baselines. We present the benefits of eye tracking features by evaluating the NER models on both individual datasets as well as in cross-domain settings.

The contents of this chapter are based on the following papers: (1) Hollenstein and Zhang (2019). Entity Recognition at First Sight: Improving NER with Eye Movement Information. *NAACL*. (2) Hollenstein et al. (2019a). Advancing NLP with Cognitive Language Processing Signals. *ArXiv*. (3) Hollenstein et al. (2020a). Towards Best Practices for Leveraging Human Language Processing Signals for Natural Language Processing. *LiNCR*.

Then, to generalize the notion that human cognitive data can be helpful for NLU models, in Section 3.3, we present improvements across a range of NLU tasks, namely, information extraction tasks, using not only eye tracking data as indirect signals of cognitive processing, but also direct measures in the form of electrical brain activity recordings. As we show in Section 3.2, and other researchers have shown on different NLU tasks (for example, Barrett et al. (2016) and Mishra et al. (2017a)), cognitive language processing data such as eye tracking features have shown improvements on single NLU tasks.

We analyze whether using such human features can show consistent improvement across NLU tasks and across cognitive data sources. Specifically, we use gaze and EEG features to augment models of named entity recognition, relation classification, and sentiment analysis. These methods significantly outperform the baselines and show the potential and current limitations of employing human language processing data for NLP. In Section 3.4, we present an extensive investigation of the benefits and limitations as well as best practices of using cognitive processing data for improving NLU models.

3.1 Background

In this section, we will shortly review the related work in this research area. We start with methods using eye tracking data to improve NLP tasks, since there is a much larger body of work and thereafter describe the few studies that have used brain activity data for NLP.

Eye Tracking

The benefits of eye movement data for machine learning have been assessed in various domains, including NLP and computer vision. As described in Chapter 2, eye trackers provide millisecond-accurate records on where humans look when they are reading, and they are becoming cheaper and more easily available by the day (San Agustin et al., 2009; Sewell and Komogortsev, 2010). Although eye tracking data is still being recorded in controlled experiment environments, this will likely change in the near future. Recent approaches have shown substantial improvements in recording gaze data while reading by using cameras of mobile devices (Gómez-Poveda and Gaudioso, 2016; Papoutsaki et al., 2016). Hence, eye tracking data will probably be more accessible and available in much larger volumes in due time, which will

facilitate the creation of sizable datasets enormously (San Agustin et al., 2009; Sewell and Komogortsev, 2010).

Tokunaga et al. (2017) recently analyzed eye tracking signals during the annotation of named entities to find effective features for NER. Their work proves that humans take into account a broad context to identify named entities, including the predicate-argument structure. This further strengthens our intuition to use eye movement information to improve existing NER systems. Going a step further, it opens the possibility for real-time entity annotation based on the reader's eye movements.

The benefit of eye movement data is backed up by extensive psycholinguistic studies. For example, when humans read a text they do not focus on every single word. The number of fixations and the fixation duration on a word depends on a number of linguistic factors (Clifton et al., 2007; Demberg and Keller, 2008). Different features even allow us to study early and late cognitive processing separately. First, readers are more likely to fixate on open-class words that are not predictable from context (Rayner, 1998). Reading patterns are a reliable indicator of syntactical categories (Barrett and Søgaard, 2015a). Second, word frequency and word familiarity influence how long readers look at a word. The frequency effect was first noted by Rayner (1977) and has been reported in various studies since, e.g., Just and Carpenter (1980) and Cop et al. (2017). Although two words may have the same frequency value, they may differ in familiarity (especially for infrequent words). Effects of word familiarity on fixation time have also been demonstrated in a number of recent studies (Juhasz and Rayner, 2003; Williams and Morris, 2004). Additionally, the positive effect of fixation information in various NLP tasks has recently been shown by Barrett et al. (2018a), where an attention mechanism is trained on fixation duration.

A range of work of using eye tracking signals to improve NLP tasks has been proposed and shows promising results. Gaze data has been used to improve tasks such as part-of-speech tagging (Barrett et al., 2016), sentiment analysis (Mishra et al., 2017a), prediction of multiword expressions (Rohanian et al., 2017), sentence compression (Klerke et al., 2016), and word embedding evaluation (Søgaard, 2016). Furthermore, gaze data has been used to regularize attention in neural architectures for NLP classification tasks (Barrett et al., 2018a).

Electroencephalography (EEG)

To the best of our knowledge, there are no applications leveraging EEG data to improve NLP tasks. However, there are good reasons to try to combine the two sources of information. EEG

could provide the missing information in eye movements to disambiguate different cognitive processes. An extended fixation duration only tells us that extended cognitive processing occurs, but not *which process*. Collecting EEG data is more expensive and time-consuming than collecting eye tracking data, which is why brain activity data is commonly less accessible. Moreover, collecting EEG data from subjects in a naturalistic reading environment is even more challenging. Hence, related work in this area is very limited.

EEG and eye tracking yield the same temporal resolution with non-invasive technologies (Sereno and Rayner, 2003). Dambacher and Kliegl (2007) found that longer fixation duration correlates with larger N400 amplitude effects. N400 is part of the normal brain response to words and other meaningful stimuli (Kutas and Federmeier, 2000). Effects of word predictability on eye movements and EEG co-registration have also been studied in serialized word representation and in natural reading (Dimigen et al., 2011). Other aspects relevant for linguistic processing can be observed in the EEG signal itself. For instance, term relevance can be associated with brain activity with significant changes in certain brain areas (Eugster et al., 2014). Differences in processing verbs and noun, concrete nouns and abstract nouns, as well as common nouns and proper nouns are also observed (Weiss and Mueller, 2003). Furthermore, there is a correspondence between computational grammar models and certain EEG effects (Hale et al., 2018).

3.2 Entity Recognition at First Sight

The field of natural language processing includes studies of tasks of different granularity and depths of semantics: from lower level tasks such as tokenization and part-of-speech tagging up to higher level tasks of information extraction such as named entity recognition, relation extraction, and semantic role labeling (Collobert et al., 2011). As NLP systems become increasingly prevalent in society, how to take advantage of information passively collected from human readers, e.g., eye movement signals, is becoming more interesting. Previous research in this area has shown promising results: Eye tracking data has been used to improve tasks such as part-of-speech tagging (Barrett et al., 2016), sentiment analysis (Mishra et al., 2017a), prediction of multiword expressions (Rohanian et al., 2017), and word embedding evaluation (Søgaard, 2016). However, most of these studies focus on either relatively lower-level tasks (e.g., part-of-speech tagging and multi-word expressions) or relatively global properties in the text (e.g., sentiment analysis). In this section, we test a hypothesis on a different level: *Can eye movement signals*

also help improve higher-level semantic tasks such as extracting information from text?

The answer to this question is not obvious. On the one hand, the quality improvement attributed to eye movement signals on lower-level tasks implies that such signals do contain linguistic information. On the other hand, it is not clear whether these signals can also provide significant improvement for tasks dealing with higher-level semantics. Moreover, even if eye movement patterns contain signals related to higher-level tasks, as implied by a recent psycholinguistic study (Tokunaga et al., 2017), noisy as these signals are, it is not straightforward whether they would help, if not hurt, the quality of the models. We provide the first study of the impact of gaze features to automatic named entity recognition from text. We test the hypothesis that eye tracking data is beneficial for entity recognition in a state-of-the-art neural named entity tagger augmented with an embedding layer of gaze features. We evaluate this hypothesis not only on the available eye tracking corpora, but also on an external benchmark dataset, for which gaze information does not exist.

Our contributions can be summarized as follows: First, we manually annotate three eye tracking corpora with named entity labels to train a neural NER system with gaze features. This collection of corpora facilitates future research in related topics. The annotations are publicly available. Second, we present a neural architecture for NER, which in addition to textual information, incorporates embedding layers to encode eye movement information. Last, we show how gaze features generalized to word types eliminate the need for recorded eye tracking data at test time. This makes the use of eye tracking data in NLP applications more feasible since recorded eye tracking data for each token in context is not required anymore at prediction time. Moreover, type-aggregated features appear to be particularly useful for cross-domain systems.

3.2.1 Eye Tracking Data

Datasets

For our experiments, we resort to three eye tracking data resources: the *Dundee corpus* (Kennedy et al., 2003), the *GECO corpus* (Cop et al., 2017) and the *ZuCo corpus* (Chapter 2). For the purpose of information extraction, it is important that the readers process longer fragments of text, i.e., complete sentences instead of single words, which is the case in all three datasets. Table 3.1 shows an overview of the domain and size of these datasets. In total, they comprise 142,441 tokens with gaze information. Table 3.1 also shows the differences in mean fixation

	Dundee	GECO	ZuCo	Total
Domain(s)	news articles	novel	movie reviews, Wikipedia articles	-
Number of sentences	2,367	5,424	700	8,491
Mean sentence length	24.75	12.65	22.12	19.84
Number of words	58,598	68,606	15,237	142,441
Unique word types	9,131	5,283	4,408	13,937
Mean word length	4.29	3.76	4.44	4.16
Fixation duration (ms)	202	214	226	214
Gaze duration (ms)	237	232	265	244.7

Table 3.1: Descriptive statistics of the eye tracking corpora, including domain, size, and mean fixation and gaze duration per token.

times between the datasets,i.e., fixation duration (the average duration of a single fixation on a word in milliseconds) and gaze duration (the average duration of all fixations on a word).

Dundee Corpus The gaze data of the Dundee corpus (Kennedy et al., 2003) was recorded with a *Dr. Bouis Oculometer Eyetracker*. The English section of this corpus comprises 58,598 tokens in 2,367 sentences. It contains eye movement information of ten native English speakers as they read the same 20 newspaper articles from *The Independent*. The text was presented to the readers on a screen five lines at a time. This data has been widely used in psycholinguistic research to analyze the reading behavior of subjects while reading sentences in context under relatively naturalistic conditions.

GECO Corpus The Ghent Eye Tracking Corpus (Cop et al., 2017) is a more recent dataset, which was created for the analysis of eye movements of monolingual and bilingual subjects during reading. The data was recorded with an *EyeLink 1000* system. The text was presented one paragraph at a time. The subjects read the entire novel *The Mysterious Affair at Styles* by Agatha Christie (1920) containing 68,606 tokens in 5,424 sentences. We use only the monolingual data recorded from the 14 native English speakers to maintain consistency across corpora.

ZuCo Corpus As presented in Chapter 2, the Zurich Cognitive Language Processing Corpus is a combined eye tracking and EEG dataset. The gaze data was also recorded with an *EyeLink*

	Dundee		GECO		ZuCo		Total	
	all	unique	all	unique	all	unique	all	unique
PERSON	732	415	1,870	108	657	446	3,259	955
ORGANIZATION	475	261	26	12	156	95	657	364
LOCATION	431	177	101	23	366	155	898	1,646
Total	**1,638**	**853**	**1,997**	**143**	**1,179**	**696**	**4,814**	**1,646**
		52%		7%		59%		34%

Table 3.2: Number and distribution of named entity annotations in all three eye tracking corpora.

1000 device. The full corpus contains 1,100 English sentences read by 12 adult native speakers. The sentences were presented at the same position on the screen one at a time. For the present work, we only use the eye movement data of the first two reading tasks of ZuCo 1.0 (700 sentences, 15,237 tokens),[1] since these tasks encouraged natural reading. The reading material included sentences from movie reviews from the Stanford Sentiment Treebank (Socher et al., 2013) and the Wikipedia dataset by Culotta et al. (2006).

For the purposes of this work, all datasets were manually annotated with named entity labels for three categories: PERSON, ORGANIZATION and LOCATION. The annotations are available online.[2] The datasets were annotated by two NLP experts. We followed the ACE Annotation Guidelines (Linguistic Data Consortium, 2005) and used the IOB tagging scheme for the labeling. All conflicts in labeling were resolved by adjudication between both annotators. An inter-annotator reliability analysis on 10,000 tokens (511 sentences) sampled from all three datasets yielded an agreement of 83.5% on the entity labels ($\kappa = 0.68$).

Table 3.2 shows the number of annotated entities in each dataset. The distribution of entities between the corpora is highly unbalanced. The Dundee and ZuCo corpora contain more heterogeneous texts and thus, have a higher ratio of unique entity occurrences. Contrarily, the GECO corpus is a homogeneous corpus consisting of a single novel, where the named entities are very repetitive.

[1] Note that the exact sentence and token counts of each eye tracking corpus may vary between the chapters and experiments of this thesis due to differing preprocessing strategies.

[2] `https://github.com/DS3Lab/ner-at-first-sight`

Eye Tracking Features

The gaze data of all three corpora was recorded for multiple readers by conducting experiments in a controlled environment using specialized equipment. It is important to consider that, while we extract the same features for all corpora, there are certainly practical aspects that differ across the datasets. The following factors are expected to influence reading: experiment procedures; text presentation; recording hardware, software and quality; sampling rates; initial calibration and filtering, as well as human factors such as head movements and lack of attention. Therefore, separate normalization for each dataset should better preserve the signal within each corpus and for the same reason the type-aggregation was computed on the normalized feature values. This is especially relevant for the type-aggregated features and the cross-corpus experiments described below.

In order to add gaze information to the neural network, we have selected as many features as available from those present in all three corpora. Previous research shows benefits in combining multiple eye tracking features of different stages of the human reading process (Tokunaga et al., 2017). The extracted features follow closely on Barrett et al. (2016). As described above, psycho-linguistic research has shown how fixation duration and probability differ between word classes and syntactic comprehension processes. Thus, the features focus on representing these nuances as broadly as possible, covering the complete reading time of a word at different stages. Table 3.3 shows the eye movement features incorporated into the experiments. We split the 17 features into 4 distinct groups (analogous to Barrett et al. (2016)), which define the different stages of the reading process:

1. *BASIC* eye tracking features capture characteristics on word-level, e.g., the number of all fixations on a word or the probability that a word will be fixated (namely, the number of subjects who fixated the word divided by the total number of subjects).

2. *EARLY* gaze measures capture lexical access and early syntactic processing and are based on the first time a word is fixated.

3. *LATE* measures reflect the late syntactic processing and general disambiguation. These features are significant for words which were fixated more than once.

4. *CONTEXT* features capture the gaze measures of the surrounding tokens. These features consider the fixation probability and duration up to two tokens to the left and right of the

Feature	Definition
BASIC	
n fixations	total number of fixations on a word w
fixation probability	the probability that a word w will be fixated
mean fixation duration	mean of all fixation durations for a word w
EARLY	
first fixation duration	duration of the first fixation on a word w
first pass duration	sum of all fixation durations during the first pass
LATE	
total fixation duration	sum of all fixation durations for a word w
n re-fixations	number of times a word w is fixated (after the first fixation)
re-read probability	the probability that a word w will be read more than once
CONTEXT	
total regression-from duration	combined duration of the regressions that began at word w
$w - k$ fixation probability	fixation probability of the k^{th} word before w
$w + k$ fixation probability	fixation probability of the k^{th} word after w
$w - k$ fixation duration	fixation duration of the k^{th} word before w
$w + k$ fixation duration	fixation duration of the k^{th} word after w

Table 3.3: Gaze features extracted from the Dundee, GECO and ZuCo corpora.

current token. Additionally, regressions starting at the current word are also considered to be meaningful for the syntactic processing of full sentences.

The eye movement measurements were averaged over all native-speaking readers of each dataset to obtain more robust estimates. The small size of eye tracking datasets often limits the potential for training data-intensive algorithms and causes overfitting in benchmark evaluation (Xu et al., 2015). It also leads to sparse samples of gaze measurements. Hence, given the limited number of observations available, we normalize the data by splitting the feature values into quantiles to avoid sparsity issues. The best results were achieved with 24 bins. This normalization is conducted separately for each corpus.

Moreover, special care had to be taken regarding tokenization, since the recorded eye tracking

data considers only whitespace separation. For example, the string *John's* would constitute a single token for eye tracking feature extraction, but would be split into *John* and *'s* for NER, with the former token holding the label PERSON and the latter no label at all. Our strategy to address this issue was to assign the same values of the gaze features of the originating token to split tokens.

Type Aggregation

Barrett and Søgaard (2015b) showed that type-level aggregation of gaze features results in larger improvements for part-of-speech tagging. Following their line of work, we also conducted experiments with type aggregation for NER. This implies that the eye tracking feature values were averaged for each word type over all occurrences in the training data. For instance, the sum of the features of all n occurrences of the token "island" are averaged over the number of occurrences n. As a result, for each corpus as well as for the aggregated corpora, a lexicon of lower-cased word types with their averaged eye tracking feature values was compiled. Thus, as input for the network, either the type-level aggregates for each individual corpus can be used or the values from the combined lexicon, which increases the number of word types with known gaze feature values.

The goal of type aggregation is twofold. First, it eliminates the requirement of eye tracking features when applying the models at test time, since the larger the lexicon, the more tokens in the unseen data receive type-aggregated eye tracking feature values. For tokens that are not in the lexicon, we assign a placeholder for unknown feature values. Second, type-aggregated features can be used on any dataset and show that improvements can be achieved with aggregated gaze data without requiring large quantities of recorded data.

3.2.2 LSTM-CRF Model

Non-linear neural networks with distributed word representations as input have become increasingly successful for any sequence labeling task in NLP (Huang et al., 2015; Chiu and Nichols, 2016; Ma and Hovy, 2016). The same applies to named entity recognition: State-of-the-art systems are combinations of neural networks such as Long Short-term Memory Networks (LSTMs; Hochreiter and Schmidhuber (1997)) or convolutional neural networks (CNNs; LeCun et al.

(1998)) and conditional random fields (CRFs; Lafferty et al. (2001)). Lample et al. (2016) developed such a neural architecture for NER, which we employ in this work and enhance with eye movement features. Their model successfully combines word-level and character-level embeddings, which we augment with embedding layers for eye tracking features.

The experiments in this work were executed using an enhanced version of the system presented by Lample et al. (2016). This hybrid approach is based on bidirectional LSTMs and CRFs and relies mainly on two sources of information: character-level and word-level representations.

LSTMs are recurrent neural networks that operate on sequential data. They take as input a sequence of vectors $(x_1, x_2, \ldots, x_n)$ and return another sequence $(h_1, h_2, \ldots, h_n)$ that represents some information about the sequence at every step in the input. LSTMs have been designed to capture long-range dependencies by using several gates that control the proportion of the input to preserve in the memory cell, and the proportion from the previous state to forget. In the equations below, the matrices W_x and W_h contain the weights of the input and recurrent connections, respectively. We use the following implementation:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \tag{3.1}$$

$$c_t = (1 - i_t) \odot c_{t-1} + i_t \odot \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{3.2}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \tag{3.3}$$

$$h_t = o_t \odot \tanh(c_t), \tag{3.4}$$

where σ is the element-wise sigmoid function, $\odot$ is the element-wise product. i_t is the input gate that controls the extent to which a information flows into the memory cell c_t and o_t is the output gate deciding over the output of c_t.

For a given sentence $(x_1, x_2, \ldots, x_n)$ containing n words, each represented as a d-dimensional vector, a forward LSTM computes a representation $\overrightarrow{h_t}$ of the left context of the sentence at every word t. Naturally, generating a representation of the right context $\overleftarrow{h_t}$ should add useful information. This can be achieved using a second backward LSTM that reads the same sequence in reverse. The pair of forward and backward LSTMs is referred to as a bidirectional LSTM (Graves and Schmidhuber, 2005). The representation of a word using this model is obtained by concatenating its left and right context representations, $h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$. This effectively includes a representation of a word in context, which is useful for numerous sequence labeling applications including named entity recognition.

Due to the hard constraints between sequences of entity labels, we model them jointly using a conditional random field (Lafferty et al., 2001). For a given input sentence $X = (x_1, x_2, \ldots, x_n)$,

we consider P to be the matrix of scores output by the biLSTM network. P is of size $n \times k$, where k is the number of distinct labels, and $P_{i,j}$ corresponds to the score of the j^{th} label of the i^{th} word in a sentence. For a sequence of predictions $y = (y_1, y_2, \ldots, y_n)$, we define its *score* to be

$$s(X, y) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i} \tag{3.5}$$

where A is a matrix of *transition scores* such that $A_{i,j}$ represents the score of a transition from the label i to label j.

To obtain a probability vector as an output, we use a *softmax* activation function, yielding

$$p(y|X) = \frac{e^{s(X,y)}}{\sum_{\tilde{y} \in Y_X} e^{s(X,\tilde{y})}}. \tag{3.6}$$

During training, we maximize the log probability of the correct label sequence:

$$\log(p(y|X)) = s(X, y) - \log\left(\sum_{\tilde{y} \in Y_X} e^{s(X,\tilde{y})}\right) \tag{3.7}$$

where Y_X represents all possible label sequences for a sentence X. While decoding, we predict the output sequence that obtains the maximum score given by:

$$y^* = \underset{\tilde{y} \in Y_X}{\operatorname{argmax}}\, s(X, \tilde{y}). \tag{3.8}$$

Training Details For the experiments, the originally proposed values for all parameters were maintained. Specifically, the bidirectional LSTMs for character-based embeddings are trained on the corpus at hand with dimensions set to 25. The lookup table for the word embeddings was initialized with the pre-trained GloVe vectors of 100 dimensions (Pennington et al., 2014). The model uses a single layer for the forward and backward LSTMs. All models were trained with a dropout rate at 0.5. Moreover, all digits were replaced with zeros.

The original model was modified to include the gaze features as additional embedding layers to the network.[3] The architecture is shown in Figure 3.1. The character-level representation, i.e., the output of a bidirectional LSTM, is concatenated with the word-level representation from a word lookup table. In the augmented model with eye tracking information, the embedding for

[3] https://github.com/glample/tagger

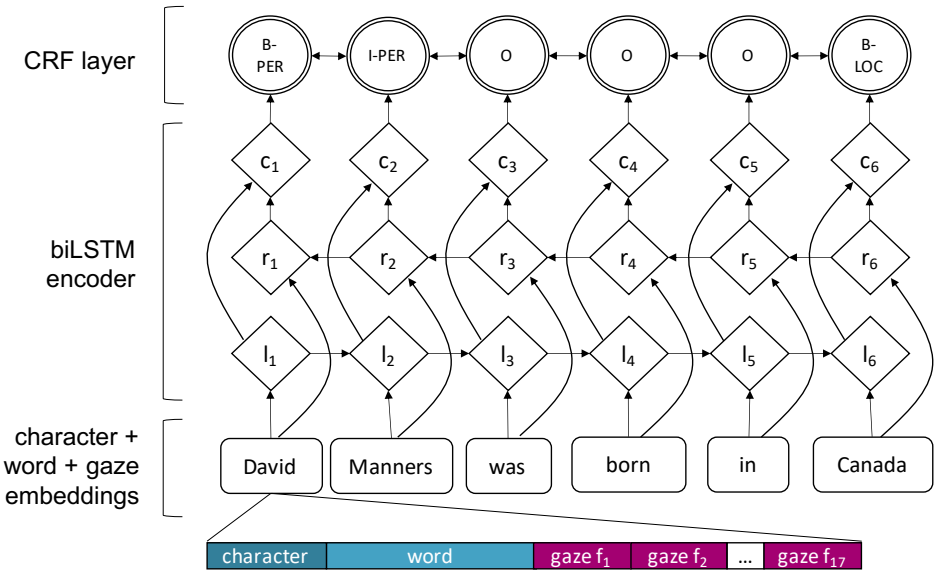


Figure 3.1: Main architecture of the named entity recognition model. Character and word embeddings concatenated with gaze features are given to a bidirectional LSTM. Here, l_i represents the word i and its left context, r_i represents the word i and its right context. Concatenating these two vectors yields a representation of the word i in its context, c_i.

each discrete gaze feature is also concatenated to the input. The dimension of the gaze feature embeddings is equal to the number of quantiles. Word length and word frequency are known to correlate and interact with gaze features (Tomanek et al., 2010), which is why we selected a base model that allows us to combine the eye tracking features with word- and character-level information.

3.2.3 Evaluation

In the following, we describe the evaluation of our augmented NER model. First, we present the results of the model trained and evaluated on single eye tracking corpora. Then, we evaluate the model on a benchmark NER dataset using type-aggregated gaze features. Additionally, we present a cross-dataset evaluation to analyze the capacities of the model in cross-domain settings. Finally, we discuss these new insights as well as the limitations of the proposed approach.

Results

We use precision, recall, and F_1-score to evaluate all NER models (see definitions in Appendix A.1). Our main finding is that our models enhanced with gaze features consistently outperform the baseline. As our baseline, we trained and evaluated the original models with the neural architecture and parameters proposed by Lample et al. (2016) on the GECO, Dundee, and ZuCo corpora and compared it to the models that were enriched with eye tracking measures. The best improvements on F_1-score over the baseline models are significant under one-sided t-tests ($p < 0.05$). All models were trained with 10-fold cross validation (80% training set, 10% development set, 10% test set) and early stopping was performed after 20 epochs of no improvement on the development set to reduce training time.

First, the performance on the individual datasets is tested, together with the performance of one combined dataset consisting of all three corpora (consisting of 142,441 tokens). In addition, we evaluate the effects of the type-aggregated features using individual type lexicons for each dataset, and combining the three type lexicons of each corpus. Finally, we experiment with cross-corpus scenarios to evaluate the potential of eye tracking features in NER for domain adaptation. Both settings were also tested on an external corpus without eye tracking features, namely, the CoNLL-2003 dataset (Tjong Kim Sang and De Meulder, 2003).

Individual Dataset Evaluation

We analyzed how augmenting the named entity recognition system with eye tracking features affects the results on the individual datasets. Table 3.4 shows the improvements achieved by adding all 17 gaze features to the neural architecture and training models on all three corpora, and on the combined dataset containing *all* sentences from the Dundee, GECO and ZuCo corpora. Noticeably, adding token-level gaze features improves the results on all datasets individually *and* combined, even on the GECO corpus, which yields a high baseline due to the homogeneity of the contained named entities (see Table 3.2).

Furthermore, Table 3.4 also presents the results of the NER models making use of the type-aggregated features instead of token-level gaze features. There are two different experiments for these type-level features: Using the features of the word types occurring in the corpus only, or using the aggregated features of all word types in the three corpora (as described above). As can be seen, the performance of the different gaze feature levels varies between datasets, but both the

Model	Precision	Recall	F_1-score
Dundee			
Baseline	79.29	78.56	78.86
With gaze	79.55	79.27	79.35
Type individual	**81.05**	**79.37**	**80.17***
Type combined	80.27	79.26	79.67
GECO			
Baseline	96.68	97.24	96.95
With gaze	**98.08**	**97.94**	**98.01***
Type individual	97.72	97.42	97.57*
Type combined	97.76	97.16	97.46*
ZuCo			
Baseline	84.52	81.66	82.92
With gaze	**86.19**	**84.28**	**85.12***
Type individual	84.21	82.61	83.30
Type combined	83.26	83.37	83.31
All			
Baseline	86.92	86.58	86.72
With gaze	88.72	89.39	89.03*
Type combined	**89.04**	**89.52**	**89.26***

Table 3.4: Precision, recall and F_1-score for all models trained on individual datasets (best results in bold; * indicates statistically significant improvements on F_1-score). *With gaze* are models trained on the original eye tracking features on token-level, *type individual* are the models trained on type-aggregated gaze features of this corpus only, while *type combined* are the models trained with type-aggregated features computed on all datasets.

original token-level features as well as the individual and combined type-level features achieve improvements over the baselines of all datasets.

To sum up, the largest improvement with eye tracking features is achieved when combining all corpora into one larger dataset, where an additional 4% is gained in F_1-score by using type-aggregated features. Evidently, a larger mixed-domain dataset benefits from the type aggre-

CoNLL-2003	Precision	Recall	F_1-score
Baseline	93.89	94.16	94.03
Type combined	**94.38**	**94.32**	**94.35***

Table 3.5: Precision, recall and F_1-score for using type-aggregated gaze features on the CoNLL-2003 dataset (* marks statistically significant improvement).

gation, while the original token-level gaze features achieve the best results on the individual datasets. Moreover, the additional gain when training on all datasets is due to the higher signal-to-noise ratio of type-aggregated features from multiple datasets.

Evaluation on CoNLL-2003 We also evaluate the type-aggregated gaze features on an external corpus with no eye movement information available. The CoNLL-2003 corpus (Tjong Kim Sang and De Meulder, 2003) has been widely used as a benchmark dataset for NER in different shared tasks. The English part of this corpus consists of Reuters news stories and contains 302,811 tokens in 22,137 sentences. We use this dataset as an additional corpus without gaze information. Only the type-aggregated features (based on the combined eye tracking corpora) are added to each word. Merely 76% of the tokens in the CoNLL-2003 corpus also appear in the eye tracking corpora described above and, thus, receive type-aggregated feature values. The rest of the tokens, those without aggregated gaze information available, receive a placeholder for the unknown feature values.

In order to avoid overfitting, we do not train on the official train/test split of the CoNLL-2003 dataset, but perform 10-fold cross validation. Applying the same experiment setting, we train the augmented NER model with gaze features on the CoNLL-2003 data and compare it to a baseline model without any eye tracking features. We achieve a minor, but nonetheless significant improvement (shown in Table 3.5), which supports the generalizability effect of the type-aggregated features on unseen data.

Cross-Domain Evaluation

In a second evaluation scenario, we test the potential of eye tracking features for NER across corpora. The goal is to leverage eye tracking features for domain adaptation. To show the robustness of our approach across domains, we train the models with token-level and type-level

		Dundee			GECO			ZuCo		
		P	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**
Dundee	Baseline				74.20	70.71	72.40	75.36	75.62	75.44
	Token				75.68	71.54	73.55*	**78.85**	74.51	77.02
	Type				**76.44**	**77.09**	**76.75***	78.33	**76.49**	**77.35**
GECO	Baseline	58.91	34.91	43.80				68.88	42.49	52.38
	Token	**59.61**	35.62	**44.53**				**69.18**	**44.22**	**53.81**
	Type	58.39	**35.99**	44.44				67.69	42.36	52.01
ZuCo	Baseline	65.85	**54.01**	59.34	83.00	**78.11**	**80.48**			
	Token	**72.62**	50.76	59.70	82.92	75.35	78.91			
	Type	69.21	53.05	**59.95**	**83.68**	74.57	78.85			

Table 3.6: Cross-corpus results: Precision (P), recall (R), and F_1-score (F_1) for all models trained on one dataset and tested on another (rows = training dataset; columns = test dataset; best results in bold; * indicates statistically significant improvements). The baseline models are trained without eye tracking features, *token* models on the original eye tracking features, and *type* are the models trained with type-aggregated features computed on all datasets.

features on 100% of corpus A, a development set of 20% of corpus B, and test on the remaining 80% of the corpus B, alternating only the development and the test set for each fold.

Table 3.6 shows the results of this cross-corpus evaluation. The impact of the eye tracking features varies between the different combinations of datasets. However, the inclusion of eye tracking features improves the results for all combinations, except for the models trained on the ZuCo corpus and tested on the GECO corpus. Presumably, this is due to the combination of the small training data size of the ZuCo corpus and the homogeneity of the named entities in the GECO corpus.

Evaluation on CoNLL-2003 Analogous to the individual dataset evaluation, we test the potential of eye tracking features in a cross-dataset scenario on an external benchmark dataset. Again, we use the CoNLL-2003 corpus for this purpose. We train a model on the Dundee, GECO *and* ZuCo data using type-aggregated gaze features and test this model on the ConLL-2003 data. Table 3.7 shows that compared to a baseline without gaze features, the results improve by 2.7% F_1-score. These results underpin our hypothesis of the possibility of generalizing eye tracking features on word type level, such that no recorded gaze data is required at test time.

CoNLL-2003	Precision	Recall	F_1-score
Baseline	72.80	56.97	63.92
Type combined	**74.56**	**60.20**	**66.61***

Table 3.7: Precision, recall and F_1-score for using type-aggregated gaze features trained on all three eye tracking datasets and tested on the CoNLL-2003 dataset (* marks statistically significant improvement).

Discussion

The models evaluated in the previous section show that eye tracking data contain valuable semantic information that can be leveraged effectively by NER systems. While the individual datasets are still limited in size, the largest improvement is observed in the models that make use of *all* the available data. At a closer look, the model leveraging gaze data yields a considerably higher increase in recall when comparing to the baselines. In addition, a class-wise analysis shows that the entity type benefiting the most from the gaze features over all models is ORGANIZATION, which is the most difficult class to predict. Figure 3.2 illustrates this with the results per class of the models trained on all three gaze corpora jointly.

In the individual dataset evaluation setting, the combined type-level feature aggregation from all datasets does not yield the best results, since each sentence in these corpora already has accurate eye tracking features on token-level. Thus, it is understandable that in this scenario the original gaze features and the gaze features aggregated only on the individual datasets result in better models. However, when evaluating the NER models in a cross-corpus scenario, the type-aggregated features lead to significant improvements.

Type aggregation evidently reduces the fine-grained nuances contained in eye tracking information and eliminates the possibility of disambiguation between homographic tokens. Nevertheless, this type of disambiguation is not crucial for named entities, which mainly consist of proper nouns and the same entities tend to appear in the same context. Especially noteworthy is the gain in the models tested on the CoNLL-2003 benchmark corpus, which shows that aggregated eye tracking features from other datasets can be applied to any unseen sentence and show improvements, even though more than 20% of the tokens have unknown gaze feature values. While the high number of unknown values is certainly a limitation of our approach, it shows at once the possibility of not requiring original gaze features at prediction time. Thus, the trained NER models can be applied robustly on unseen data.

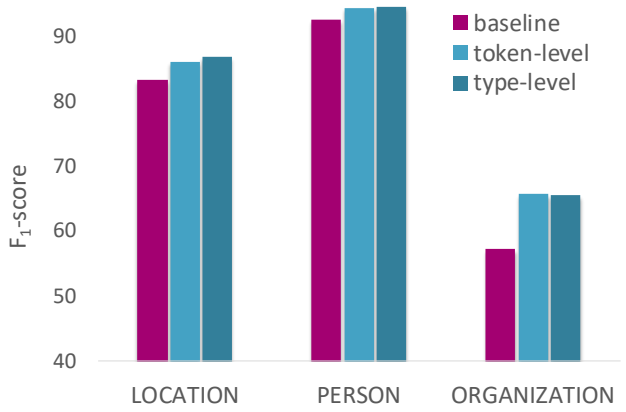


Figure 3.2: Token-level and type-level results per class for the models trained on all gaze datasets combined.

3.3 Generalizing Across Tasks and Cognitive Signals

When reading, humans process language "automatically" without reflecting on each step. Humans string words together into sentences, understand the meaning of spoken and written ideas, and process language without thinking about how the underlying cognitive process happens. This process generates cognitive signals that could potentially facilitate natural language processing tasks.

In recent years, collecting these signals has become increasingly easy and less expensive (Papoutsaki et al., 2016); as a result, using cognitive features to improve NLP tasks has become more popular. For example, researchers have proposed a range of work that uses eye tracking or gaze signals to improve part-of-speech tagging (Barrett et al., 2016), sentiment analysis (Mishra et al., 2017a), or named entity recognition as described previously, among other tasks. Moreover, these signals have been used successfully to regularize attention in neural networks for NLP (Barrett et al., 2018a).

However, most previous work leverages only eye tracking data, presumably because it is the most accessible form of cognitive language processing signal. In addition, most state-of-the-art work focused on improving a single task with a single type of cognitive signal. An interesting question is whether *cognitive processing signals bring consistent improvements across modalities (e.g., eye tracking and/or EEG) and across various NLP tasks?* And if so, *does the combination of different sources of cognitive signals bring incremental improvements?*

In this section, we aim at shedding light on these questions. We present, to the best of our knowledge, the first comprehensive study to analyze the benefits and limitations of using cognitive language processing signals to improve NLP across multiple tasks and modalities (types of signals). Specifically, we go beyond state-of-the-art in two ways:

- **Multiple Signals:** We consider both eye tracking and electroencephalography (EEG) data as examples of cognitive language processing data. Eye tracking records the readers gaze positions on the screen and serves as an *indirect* measure of the cognitive reading process. EEG records electrical brain activity along the scalp and is a more *direct* measure of physiological processes, including language processing. This is also the first application leveraging EEG data to improve NLP tasks.
- **Multiple Tasks:** We construct named entity recognition, relation classification, and sentiment analysis models with gaze and EEG features. We analyze three methods of adding these cognitive signals to machine learning architectures for NLP. First, we simply add the features to existing systems (Section 3.3.2). Second, we show how these features can be generalized so that recorded data is not required at test data (Section 3.3.3). And third, in a multi-task setting, we learn gaze and EEG features as auxiliary tasks to aid the main NLP task (Section 3.3.4).

In summary, the most important insights gained from the experiments in this section are the following. (1) Using cognitive features shows consistent improvements over a range of NLP tasks even without large amounts of recorded cognitive signals. (2) While integrating gaze or EEG signals separately significantly outperforms the baselines, the combination of both does not further improve the results. (3) We identify multiple directions for future research: *How can cognitive signals, such as EEG data, be preprocessed and de-noised more efficiently for NLP tasks? How can cognitive features of different sources be combined more effectively for natural language processing?* The code for all experiments presented in this chapter is available to provide a foundation for future work to better understand these questions.[4]

3.3.1 Data

The Zurich Cognitive Language Processing Corpus (ZuCo) is the main data source of this work. It is the first freely available dataset of simultaneous eye tracking and EEG recordings of natural

[4] https://github.com/DS3Lab/zuco-nlp/

sentence reading. This corpus includes recordings of 12 adult, native speakers reading approximately 1,100 English sentences. In this work, we used only ZuCo 1.0, since ZuCo 2.0 was not yet available at the time.

As described in Chapter 2, the corpus contains both natural reading and task-solving reading paradigms. For this work, we use the first two reading paradigms of ZuCo 1.0, during which the subjects read naturally at their own speed and without any specific task other than answering some control questions testing their reading comprehension. The first paradigm includes 300 sentences (7,737 tokes) from Wikipedia articles (Culotta et al., 2006) that contain semantic relations such as *employer*, *award* and *job_title*. The second paradigm contains 400 positive, negative, and neutral sentences (8,138 tokens) from the Stanford Sentiment Treebank (Socher et al., 2013), to analyze the elicitation of emotions and opinions during reading.

Eye Tracking Features

As described in Chapter 2, ZuCo readily provides 5 eye tracking features: number of fixations (NFIX), the number of all fixations landing on a word; first fixation duration (FFD), the duration of the first fixation on the current word; total reading time (TRT), the sum of all fixation durations on the current word; gaze duration (GD), the sum of all fixations on the current word in the first-pass reading before the eye moves out of the word; and go-past time (GPT), the sum of all fixations prior to progressing to the right of the current word, including regressions to previous words that originated from the current word. Fixations shorter than 100 ms were excluded, since these are unlikely to reflect language processing (Sereno and Rayner, 2003). To increase the robustness of the signal, the eye tracking features are averaged over all subjects.

EEG Features

Since eye tracking and EEG were recorded simultaneously, we were able to extract word-level EEG features. During the preprocessing of ZuCo, 23 electrodes in the outermost circumference (chin and neck) were used to detect muscular artifacts and were removed for subsequent analyses. Thus, each EEG feature, corresponding to the duration of a specific fixation, contains 105 electrode values. The EEG signal is split into 8 frequency bands, which are fixed ranges of wave frequencies and amplitudes over a time scale: $theta_1$ (4-6 Hz), $theta_2$ (6.5-8 Hz), $alpha_1$ (8.5-10 Hz), $alpha_2$ (10.5-13 Hz), $beta_1$, (13.5-18 Hz) beta_2 (18.5-30 Hz), $gamma_1$ (30.5-40 Hz) and

$gamma_2$ (40-49.5 Hz). These frequency ranges are known to correlate with certain cognitive functions. For instance, theta activity reflects cognitive control and working memory (Williams et al., 2019), alpha activity has been related to attentiveness (Klimesch, 2012), gamma-band activity has been used to detect emotions (Li and Lu, 2009) and beta frequencies affect decisions regarding relevance (Eugster et al., 2014). Even though the variability between subjects is much higher in the EEG signal, we also average all features over all subjects.

3.3.2 Tasks

To thoroughly evaluate the potential of gaze and brain activity data, we perform experiments on the three information extraction tasks described in this section. Current state-of-the-art systems are used for all tasks and different combinations of cognitive features are evaluated.

Named Entity Recognition

As we showed in Section 3.2, the performance of named entity recognition (NER) systems can successfully be improved with eye tracking features. However, this has not been explored for EEG signals. We use the same LSTM-CRF architecture for NER by Lample et al. (2016) that we described previously. This model successfully combines word-level and character-level embeddings, which we augment with embedding layers for gaze and/or EEG features. Word length and frequency are known to correlate and interact with gaze features (e.g., Just and Carpenter (1980); Rayner (1977)), which is why we selected a base model that allows us to combine the cognitive features with word-level and character-level information. We use the named entity annotations presented in Section 3.2.

Features For this task, we used same the 17 gaze features proposed in Section 3.2. These features include relevant information from early and late word processing as well as context features from the surrounding words. We extracted 8 word-level EEG features, one for each frequency band. The neural architecture of this system does not allow for raw normalized EEG and gaze features as is the case for relation classification and sentiment analysis. The feature values were averaged over the 105 electrode values. These features are mapped to the duration of the gaze features. Thus, in the experiments we tested EEG features during the total reading time

of the words and EEG features merely during the first fixation. The latter yielded better results. The gaze and EEG feature values (originally in milliseconds for eye tracking and microvolts for EEG) were normalized and concatenated to the character and word embeddings as one-hot vectors.

Experiments All models were trained on both ZuCo paradigms described above (15,875 tokens) with 10-fold cross validation (80% training, 10% development, 10% test) and early stopping was performed after 20 epochs of no improvement on the development set to reduce training time. For the experiments, the default values for all parameters were maintained. The word embeddings were initialized with the pre-trained GloVe vectors of 100 dimensions (Pennington et al., 2014) and the character-based embeddings were trained on the corpus at hand (25 dimensions). The model uses a single layer for the forward and backward LSTMs. All models were trained with a dropout rate of 0.5.

Relation Classification

The second information extraction task we analyze is relation classification. Relation classification is the task of identifying the semantic relation holding between two entities in text. As a state-of-the-art relation classification method, we use the winning system from SemEval 2018 (Rotsztejn et al., 2018), which combines convolutional and recurrent neural networks to leverage the best architecture for different sentence lengths. We consider the following 11 relation types: *award, employer, education, founder, visited, wife, political-affiliation, nationality, job-title, birth-place* and *death-place*. We use the annotations provided by Culotta et al. (2006).

Features For this task, we employed the 5 gaze features on word-level provided in the ZuCo data: number of fixations, first fixation duration, total reading time, gaze duration, and go-past time. The eye tracking feature values were normalized over all occurrences in the corpus. The EEG features were extracted by averaging the 105 electrode values over all fixations for each word and then normalized. All word features in a sentence were concatenated and finally padded to the maximum sentence length. The eye tracking and/or EEG feature vectors were appended to the word embeddings.

Experiments We performed 5-fold cross validation over 566 samples (sentences can include more than one relation type). We split the data into 80% training data and 20% test data. Due to

the small size of the dataset, we used the same preprocessing steps and parameters as proposed by the SemEval 2018 system. The word embeddings were initialized with the pre-trained GloVe vectors of 300 dimensions.

Sentiment Analysis

The third NLU task we choose for this work is sentiment analysis. The objective of sentiment analysis is to interpret subjective information in text. More specifically, we define sentiment analysis as a sentence-level classification task. Based on the analysis by Barnes et al. (2017), we implemented a bidirectional LSTM with an attention layer for the classification of sentence-level sentiment labels.

Features Analogous to the relation classification, the 5 word-level eye tracking features were normalized and concatenated before being appended to the sentence embeddings. The raw EEG data (105 electrode values per word) were averaged and normalized.

Experiments We perform 10-fold cross-validation over the 400 sentences with available sentiment labels from ZuCo (123 neutral, 137 negative, and 140 positive sentences). We test ternary classification as well as binary classification. For the latter, we remove all neutral sentences from the training data. Word embeddings were initialized with pre-trained vectors of 300 dimensions (Mikolov et al., 2013a). All models are trained for 10 epochs with batch sizes of 32. The initial learning rate is set to 0.001 and it was halved every 3 passes for binary classification and every 10 passes for ternary classification (due to the larger training set).

3.3.3 Evaluation

For each information extraction task described in the previous section we trained baseline models, models augmented with gaze features, with EEG features, and with both. All baseline models were trained solely on textual information (i.e., word embeddings without any gaze or EEG features). We trained single-subject models and models in which the feature values are averaged over all subjects.

	NER			RelClass		
	Precision	**Recall**	**F_1-score**	**Precision**	**Recall**	**F_1-score**
Baseline	84.5	81.7	82.9	62.6	56.6	57.7
Gaze	86.2	84.3	**85.1****	65.1	61.9	62.0**
EEG	86.7	81.5	83.9*	68.3	64.8	**65.1****
Gaze+EEG	85.1	83.2	84.0**	66.3	59.3	60.8**

Table 3.8: Precision, recall and F_1-score for the NER and relation classification tasks augmented with gaze features, EEG features, and both. Significance is indicated with the asterisks: * = p<0.01, ** = p<0.0008 (Bonferroni correction).

	Sentiment (2)			Sentiment (3)		
	Precision	**Recall**	**F_1-score**	**Precision**	**Recall**	**F_1-score**
Baseline	82.5	82.5	82.5	57.1	57.6	57.2
Gaze	84.7	84.6	**84.6****	61.4	61.7	**61.5****
EEG	83.6	83.6	83.6**	60.5	60.2	60.3**
Gaze+EEG	84.3	84.3	84.3**	59.8	60.0	59.8**

Table 3.9: Precision, recall and F_1-score for binary and ternary sentiment analysis tasks augmented with gaze features, EEG features, and both. Significance is indicated with the asterisks: * = p<0.01, ** = p<0.0008 (Bonferroni correction).

The results of the averaged models are shown in Tables 3.8 and 3.9. We observe consistent improvements over the baselines for all tasks when augmented with cognitive features. The models with gaze features, EEG features, and the combination thereof all outperform the baseline. Notably, while the combinations of gaze and EEG features also outperform the baseline, they do not improve over using gaze or EEG individually.

We perform statistical significance testing using permutation (as described in Dror et al. (2018)) over all tasks. In addition, we apply the conservative Bonferroni correction (Bonferroni, 1936) for multiple hypotheses, where the global null hypothesis is rejected if $p < \alpha/N$, where N is the number of hypotheses (Dror et al., 2017). In our setting, $\alpha = 0.01$ and $N = 12$, accounting for the combination of the 4 tasks and 3 configurations (EEG, gaze, EEG+gaze). The improvements in 11 configurations out of 12 are also statistically significant under the Bonferroni correction. Bonferroni's method does not make any assumptions about the dependencies between the par-

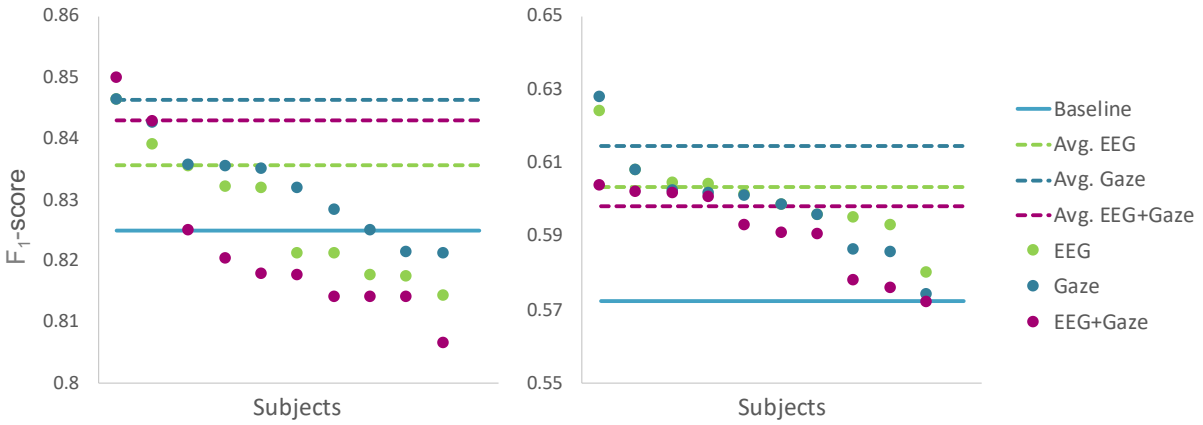


Figure 3.3: Comparison of single-subject models and features averaged over all subject for both binary sentiment classification (left) and ternary sentiment classification (right). Each dot represents a single subject model, each line an averaged feature model. Note that the best-performing subject for gaze is not necessarily the same subject as for the best EEG model.

ticipating datasets. Hence, it is applicable for the NLP tasks described in this work, due to the dependence between the datasets used for training. Despite the limited amount of data, this result suggests that augmenting NLP systems with cognitive features is a generalizable approach.

Subject Analysis

We evaluate the single-subject models to test the robustness of averaging the feature values over all readers. By the example of binary and ternary sentiment analysis, Figure 3.3 depicts the variability of the results between the subjects. In contrast to the averaged models, the best subject for binary sentiment classification reaches an F_1-score of 85% with the combination of gaze and EEG data. Moreover, it shows how the averaged models perform almost as well as the best subject. Note that the best-performing subject for gaze is not necessarily the same subject as for the best EEG model. We also trained models that only take into account the feature values of the five best subjects. However, when averaging over *all* subjects, the signal-to-noise ratio is higher and provides better results than training on the best five subjects only. While previous research had shown the same effect when using eye tracking data from multiple subjects in NLP (Rohanian et al., 2017; Yaneva et al., 2018), this had not yet been shown for EEG data.

	NER	RelClass	Sentiment (2)	Sentiment (3)
Corpus	CoNLL-2003	WIKIPEDIA	SST	SST
Tokens	302,811	32,953	165,165	202,125
Sentences	22,137	1,794	9,612	11,853
Unknown tokens	41.09%	30.31%	26.02%	25.96%
Baseline	94.02	76.94	82.01	57.13
Gaze	94.41**	**77.85****	81.64	**57.48****
EEG	94.58**	76.40	80.07	54.27
Gaze + EEG	**94.63****	77.01	79.74	54.80

Table 3.10: The top part shows the size of the datasets used for the type-aggregation experiments, including the percentage of unknown tokens, i.e., tokens not in the lexicon of aggregated type features. The bottom part shows F_1-scores of type aggregation on external benchmark corpora. Significance is indicated with the asterisks: * = p<0.01, ** = p<0.0008 (Bonferroni correction).

NO REAL-TIME RECORDED DATA REQUIRED

While adding these cognitive features to a system shows the potential of this type of data, it is not very practical if real-time recordings of EEG and/or eye tracking are required at prediction time. Following Barrett et al. (2016), we evaluate feature aggregation on word-type level. This means that all cognitive features are averaged over the word occurrences. As a result, a lexicon of lower-cased word types with their averaged gaze and EEG feature values was compiled. Words in the training data, as well as in the test set, are assigned these features if the words occur in the type-aggregated lexicon, or receive unknown feature values otherwise. Thus, recorded human data is not required at test time.

We evaluate the concept of type aggregation on the tasks described above. We choose 3 benchmark datasets and add the aggregated EEG and/or eye tracking features to words occurring in ZuCo. For NER, we use the CoNLL-2003 corpus (Tjong Kim Sang and De Meulder, 2003); for relation classification, we use the full Wikipedia dataset provided by (Culotta et al., 2006); and for sentiment analysis, we use the Stanford Sentiment Treebank (SST). The same experiment settings as above were applied here. To avoid overfitting, we did not use the official train/test splits but performed cross-validation. Table 3.10 shows the details about these datasets and the results. We can observe a consistent improvement using type-aggregated gaze features. How-

ever, the effect of type-aggregated EEG features is mixed. Type aggregation shows not only that recorded gaze or EEG data is not necessary at test time, but also that improvements can be achieved with human data without requiring large quantities of recorded data.

3.3.4 Multi-Task Learning

We further investigate multi-task learning (MTL) as an additional machine learning strategy to benefit from cognitive features. The intuition behind MTL is that training signals of one task, the auxiliary task, improves the performance of the main task by sharing information throughout the training process. In our case, we learn gaze and EEG features as auxiliary tasks to improve the main NLP task.

In previous work, MTL has been used successfully for sequence labeling tasks (Bingel and Søgaard, 2017) due to some compelling benefits, including its potential to efficiently regularize models and to reduce the need for labeled data. Moreover, gaze duration has been predicted as an auxiliary task to improve sentence compression (Klerke et al., 2016), and to better predict the readability of texts (González-Garduño and Søgaard, 2018). To the best of our knowledge, EEG features have not been used in MTL to improve NLP tasks.

In multi-task learning, it is important that the tasks that are learned simultaneously are related to a certain extent (Caruana, 1997; Collobert et al., 2011). Assuming that the cognitive processes in the human brain during reading are related, there should be a gain from training on gaze and EEG data when learning to extract information from text. Thus, we assess the hypothesis that MTL might also be useful in our scenario.

Experiments We used the *Sluice* networks (Ruder et al., 2017), where the network learns to which extent the layers are shared between tasks. We re-formulated the sentiment analysis as sequence labeling tasks on phrase level. For binary sentiment analysis, the classes NEUTRAL and NOT-NEUTRAL were predicted. We did not have to modify the named entity recognition task and the relation classification was not tested since only sentence level labels are available.

We ran 5-fold cross-validation for all experiments over the same data as described in Section 3.3.1. As our baselines we used single-task learning and learning word frequency as an auxiliary task to an NLP task. Word frequencies were extracted from the British National Corpus (Kilgarriff, 1995). The experiments ran with the default settings recommended by (Ruder et al.,

Main task	Auxiliary task(s)	Accuracy
NER	-	87.34
	Word frequency	91.29
	FFD	87.34
	FFD, word frequency	**91.87**
	EEG_a	87.31
	EEG_a, word frequency	91.79
Sentiment binary	-	60.99
	Word frequency	61.15
	TRT	61.31
	TRT, word frequency	61.13
	EEG_b	61.01
	EEG_b, word frequency	**61.56**
Sentiment ternary	-	61.03
	Word frequency	61.02
	FFD	61.05
	FFD, word frequency	61.10
	EEG_t	61.05
	EEG_t, word frequency	**61.17**

Table 3.11: Results of the multi-task learning experiments on NER, binary and ternary sentiment analysis.

2017). In accordance to their results, the *Sluice* networks yielded consistently higher results than hard parameter sharing.

As a main task, the network learned to predict NER, binary or ternary sentiment labels. As auxiliary tasks, the network learned a single gaze or EEG feature. We used five eye tracking features: number of fixations (nFIX), mean fixation duration (MFD), first fixation duration (FFD), total reading time (TRT), and fixation probability (FIXP). Additionally, we tested four EEG features, one for each combined frequency band: EEG_t (i.e., the average values of $theta_1$ and $theta_2$), EEG_a, EEG_b, EEG_g. The features were discretized and binned.

Results Table 3.11 shows the results of these experiments. Note that only the best feature combinations are included in the table. Learning word frequency as an auxiliary task is a strong

baseline. Learning gaze and EEG features as auxiliary tasks does not improve the performance over the single-task baseline for NER, whereas the improvements for sentiment analysis are minimal. Learning two auxiliary tasks, a gaze of EEG feature *and* word frequency in parallel yields modest improvements over the frequency baseline. Adding further auxiliary tasks with additional gaze or EEG features did not yield better results. Moreover, the combination of learning gaze and brain activity features did also not bring further improvements.

As we know that gaze and frequency band EEG features represent different cognitive processes involved in reading, our main and auxiliary tasks should in fact be related. However, it seems that the noise-to-signal ratio in the EEG features is too high to achieve significant results. As stated by González-Garduño and Søgaard (2018), it is important to establish whether the same feature representation can yield good results for all tasks independently. To gain further insights into these results, we analyze how well these human features can be learned.

Learning Cognitive Features

Using the same experiment setting as for the above-described MTL experiments, we first trained single-task baselines for each of the gaze and EEG features. Then, we trained each gaze feature in 3 MTL settings: (1) word frequency as an auxiliary task, (2) the remaining gaze features as parallel auxiliary tasks, and (3) the EEG features as parallel auxiliary tasks. The same applies to EEG features as main tasks. The results in Table 3.12 show that gaze features have far higher baselines than EEG features. Presumably, EEG is harder to learn because it has larger variance in the data. Moreover, while the eye tracking data is limited to the visual component of the cognitive processes, EEG data additionally contains a motor component and a semantic component during the reading process.

Learning word frequency as an auxiliary task considerably helps all gaze and EEG features. This might attributed to the data distribution of the frequency feature (Alonso and Plank, 2017). The known correlation between eye tracking and word frequency (Rayner and Duffy, 1986) is clearly beneficial for learning gaze features. Moreover, a frequency effect can also be found in *early* EEG signals, i.e., during the first 200ms of reading a word (Hauk and Pulvermüller, 2004).

Benefits & Limitations

In accordance with previous work (e.g., Barrett et al. (2016); Mishra and Bhattacharyya (2018)), we showed consistent improvements when using gaze data in a range of information extraction

	Gaze features					EEG features			
	nFIX	MFD	FFD	TRT	FIXP	EEG_t	EEG_a	EEG_b	EEG_g
-	64.14	84.60	55.21	65.04	46.66	40.67	36.14	39.50	30.48
Frequency	71.01	84.64	**63.68**	71.99	**56.34**	53.36	**49.75**	**52.79**	**41.34**
Gaze	**71.34**	**84.78**	63.60	**72.20**	55.77	**53.53**	49.38	52.58	40.95
EEG	71.15	84.64	62.10	72.03	55.63	53.47	46.77	52.54	37.27

Table 3.12: Accuracy of learning cognitive features in an MTL setting. Columns = main tasks, rows = auxiliary tasks.

tasks, with recorded token-level features and with type-aggregated features on benchmark corpora. The patterns in the results are less consistent when enhancing NLP methods with EEG signals. While we can still show significant improvements over the baseline models, in general, the models leveraging EEG features yield lower performance than the ones with gaze features. A plausible explanation for this is that the combination of gaze and EEG features decreases the signal-to-noise ratio even more than when only one type of cognitive data is used. Another interpretation is that the eye tracking and EEG signal contain information that is (too) similar. Thus, the combination does not yield better results.

Consequently, there are some open questions: How can EEG signals be preprocessed and denoised more efficiently for NLP tasks? How can EEG and eye tracking (and other cognitive processing signals or fortuitous data (Plank, 2016)) be combined more effectively to improve NLP applications? We address some of these open challenges in Chapter 4.

The models leveraging type-aggregated features show that improvements can be achieved without requiring large amounts of recorded data and provide evidence that this type of data can be generalized on word type level. Although these results indicate that large amounts of recorded data are not necessary for performance gains, one of the limitations of this work is the effort of collecting cognitive processing signals from humans. However, webcam-based eye trackers (e.g., Papoutsaki et al. (2016)) and commercially available EEG devices (e.g Stytsenko et al. (2011)) are becoming more accurate and user-friendly.

The multi-task learning experiments provide insights into the correlation of learning NLP tasks together with word frequency and cognitive features. While the results are not as promising as our main experiments, it reveals the qualities of the individual gaze and EEG features.

For future work, a possible approach to combine the potential of exceptionally good single-

subject models and multi-task learning, would be to learn gaze and/or EEG features from multiple subjects at the same time. This has been shown to improve accuracy on brain-computer interface tasks and helps to further reduce the variability between subjects (Panagopoulos, 2017). One of the challenges of NLP is to learn as much as possible from limited resources. Using cognitive language processing data may allow us take a step towards meta-reasoning, the process of discovering the cognitive processes that are used to tackle a task in the human brain (Griffiths et al., 2019), and in turn be able to improve NLP.

3.4 Discussion

When we want to use cognitive signals to improve our computational models, we are facing multiple modeling decisions. In this section, we discuss the advantages and disadvantages of eye tracking and EEG signals, the aspects to consider when choosing a dataset, as well as which features can be extracted from the cognitive data, and finally, how they can be included in machine learning models and how these should be evaluated. The decision of which type of signal to work with and which dataset to use depend strongly on the type of research questions that we would like to address. In this section, we provide some guidelines on how to approach these decisions based on the lessons learned from the work presented in this chapter.

Choosing the Type of Cognitive Signal

Eye tracking, as an indirect metric of cognitive load, has numerous advantages. It is an accessible method to record millisecond-accurate eye movements and has successfully been leveraged to improve a wide range of NLP tasks. While the improvements on precision and recall are modest, they are consistent across tasks. The impressive body of psycholinguistic research, a range of established metrics, and the intuitive linking from features to words speak in favour of using eye tracking for NLP.

EEG is also a recording technique with very high temporal resolution (i.e., resulting in multiple samples per second). However, as the electrodes measure electrical activity at the surface of the brain, it is difficult to know exactly in which brain region the signal originated. EEG signals have been used frequently for classification in brain-computer-interfaces (e.g., classifying text difficulty for speech recognition (Chen et al., 2012)), but have rarely been used to improve NLP

tasks. We presented the first such approach in Section 3.3. In Chapter 4 we analyze some open questions regarding which EEG features are most appropriate, noting that not much EEG data from naturalistic reading is yet openly available.

Evidently, human language processing recordings are very noisy. Therefore, if possible, it is advisable to work with multiple datasets of the same modality, or to work with multiple modalities to achieve more robust results. It is insightful to run experiments on multiple cognitive datasets of the same modality. This ensures that the NLP models are not merely picking up on the noise in the cognitive data, but actually learning from language processing specific signals.

Selecting a Dataset of Cognitive Processing Signals

Datasets of human language processing signals should be chosen based on the research question. It is important to decide whether controlled experiments with clearly distinguishable conditions are required, for instance, if infrequent linguistic phenomena are of interest, or if natural stimuli are favorable to analyze real-world language (Hamilton and Huth, 2018). Due to the different scopes in experimental research and NLP, it is seldom possible to directly draw conclusions concerning features from these studies to NLP. Speaking in broad terms, psycholinguistic and neurolinguistic studies provide evidence of human cognitive processing of text or speech primarily through controlled experiments. The experiments, as well as the textual stimuli, are carefully designed in order to isolate a specific cognitive process. Data-driven NLP works towards enabling computers to understand and manipulate naturally occurring human language through machine learning models based on large corpora. The phenomena that NLP models aim to model are typically much broader and less clearly defined than what is examined in psycholinguistic studies.

As described in Chapter 2, it has become more common to implement naturalistic reading experiments (Hamilton and Huth, 2018). This allows subjects to read at their own speed and results in different reading times between subjects, which calls for more elaborate pre-processing. In addition to the more natural setting, a big advantage is the possibility to study linguistic phenomena on different levels (e.g., phonemes, syllables, words, phrases, sentences, and discourse), which unfold at different timescales in the same naturalistic stimulus such as a story. Moreover, naturalistic experimental designs, which use language within the rich context of stories, audiobooks, and dialogues, produce results which are more easily generalizable to everyday language use (Kandylaki and Bornkessel-Schlesewsky, 2019). Since the generalizability of results is one

of the main objectives in experimental science, the potential importance of increased ecological validity in naturalistic experiment paradigms is undeniable.

An example for the use of continuous naturalistic stimuli is the ZuCo dataset presented in Chapter 2. We recorded eye tracking and EEG signals of participants silently reading full real-world sentences. In Broderick et al. (2018) and Shain et al. (2019) subjects listen to full stories during EEG and fMRI recordings, respectively. In addition to the studies mentioned in this section, a collection of openly available cognitive datasets useful for NLP in various languages can be found online.[5]

Extracting Cognitive Features

NLP studies that leverage human gaze signals from reading mostly use a broad range of established features, encompassing both early and late measures of cognitive processing. These features are then used in machine learning systems to learn patterns. Barrett et al. (2016) use 22 features for part-of-speech induction and Strzyz et al. (2019) use 12 features for dependency parsing. Similarly, in the experiments described in Section 3.2, we use 17 features for named entity recognition. Studies that systematically test different combinations of features generally reveal that using a broad range of established features, such as first, mean and total fixation duration, yield the largest improvements (Barrett et al., 2016; Yaneva et al., 2018; Rohanian et al., 2017).

In this chapter, we used eye movement and EEG features to improve named entity recognition, relation classification, and sentiment classification. We showed that averaging over ten skilled native readers is able to diminish the noise and variability between subjects, to the extent where the average worked almost as well as the best individual reader, for both gaze and EEG models. Most studies combine linguistic features with gaze features (e.g., Rohanian et al. (2017) and Yaneva et al. (2018)). Further, Barrett et al. (2016) use word frequency and word length features in combination with eye tracking features, because the two properties explain much of the variance in fixation duration (Just and Carpenter, 1980; Levy, 2008). In our experiments as well as in Barrett et al. (2018b), gaze features are combined with pre-trained word embeddings to improve performance.

[5] https://github.com/norahollenstein/cognitiveNLP-dataCollection

All these works, however, rely on rather heavy feature engineering. Contrariwise, these features can also be predicted from text: Hahn and Keller (2016) presented an unsupervised neural model of human reading by predicting the fixations within sentences. Similarly, Matthies and Søgaard (2013) predict skipping probabilities across multiple readers. Moreover, Singh et al. (2016) introduced a method where eye movements are learned in order to alleviate the need to get the task data annotated with eye movements. We take a closer look at the possibility of predicting eye tracking features in Chapter 5.

In some studies, averages of gaze features over word types have been used to alleviate the need of having gaze data at test time, and even achieved better results than token-level features (Barrett et al., 2016). Klerke and Plank (2019) analyzed this in detail for part-of-speech tagging and found that content words are especially sensitive to type-level gaze features. We were able to achieve similar improvements with aggregating eye tracking features for NER, however, the results were not equally good when aggregating EEG features. Type-aggregation eliminates the need of recorded data at test time. However, the results are more promising for eye tracking data than for brain activity.

Multilingual Neurolinguistics

The majority of research in NLP, as well as most of the available cognitive data sources, are in English. However, it is well known that language processing between native and foreign language speakers differs in the active brain regions (Perani et al., 1996). Moreover, second language learners exhibit different reading patterns than native speakers (Dussias, 2010). Eye tracking and fMRI studies on bilingualism suggest that, although the same general structures are active for both languages, differences within these general structures are present across languages and across levels of processing (Marian et al., 2003; Dehghani et al., 2017). Further, there are even differences in the processing of dialects and standard variations, e.g., Lundquist and Vangsnes (2018) for Norwegian dialects and Stocker and Hartmann (2019) for variations of German. Hence, it is not only important to take language-specific aspects into account in the NLP methods, but it is crucial to account for these differences in human language processing. It remains an open question how many of the experiments in this thesis would generalize to other languages. We take a first step in this direction of research in Chapter 5, where we learn eye tracking features recorded from reading texts in various languages.

3.5 Summary

We presented the first study of augmenting a NER system with eye tracking information. Our results highlight the benefits of leveraging cognitive cues such as eye movements to improve entity recognition models. The manually annotated named entity labels for the three eye tracking corpora are freely available. We augmented a neural NER architecture with gaze features. Experiments were performed using a wide range of features relevant to the human reading process and the results show significant improvements over the baseline for all corpora individually. In addition, the type-aggregated gaze features are effective in cross-domain settings, even on an external benchmark corpus. The results of these type-aggregated features are a step towards leveraging eye tracking data for information extraction at training time, without requiring real-time recorded eye tracking data at prediction time.

Next, we presented an extensive study of improving NLU tasks with eye tracking and electroencephalography data as instances of cognitive processing signals. We showed how adding gaze and/or EEG features to a range of information extraction tasks, namely, named entity recognition, relation classification, and sentiment analysis, yields significant improvements over the baselines. Moreover, we showed how these features can be generalized at word type-level so that no recorded data is required during prediction time. Finally, we explored a multi-task learning setting to simultaneously learn NLU tasks and cognitive features.

In conclusion, the eye tracking and EEG signals recorded from humans reading text, even though noisy and available in limited amounts, show great potential in improving NLU tasks and facilitate insights into language processing which can be applied to NLU, but need to be investigated in more depth. Therefore, in the following two chapters we address the largest challenges we encountered. In Chapter 4 we analyze how to decode EEG signals for the multi-modal learning of NLU tasks, and in Chapter 5, we learn to predict cognitive features so that we can tackle the problem of limited training data available for such augmented cognitive NLU models.

Chapter 4

Decoding Brain Activity for NLP

The results of the work presented in Chapter 3 clearly show the potential of leveraging human language processing data for improving NLP models. However, it also uncovers open challenges. Is still an open question as to how electrical brain activity data is best processed for machine-learning based language understanding tasks. Moreover, until recently, human behavioral data from reading has mainly been of interest to researchers to understand human cognition. fMRI and eye tracking have been used for augmenting natural language processing models; while signals such as electroencephalography (EEG) are largely unexplored in this context.

In Section 4.1, we first describe the motivation for conducting a large-scale study of systematically analyzing the potentials of EEG brain activity data for augmenting NLP tasks, with special focus on which features of the signal are most beneficial. To better isolate the effect of the cognitive features, we move from feature-augmented models to a late-fusion multi-modal architecture for NLP classification tasks. The architecture we present in Section 4.2 is more flexible in terms of adding various types of cognitive data and is inspired by related work in multi-modal learning. We find that filtering the EEG signals into frequency bands is more beneficial than using the full signal (Section 4.3). Furthermore, for a range of word embeddings types, EEG data improves binary and ternary sentiment classification and outperforms baselines with and without eye tracking data. For more complex tasks such as relation detection, further research is needed.

The contents of this chapter are largely based on the following publication: Hollenstein et al. (2020c), A Large-Scale Study of Decoding EEG Brain Activity for Multi-Modal Natural Language Processing, *Under Review*.

Finally, for Glove and BERT representations, EEG data shows to be particularly promising for limited training data.

4.1 Background

In recent years, natural language processing (NLP) researchers have increasingly leveraged human cognitive language processing signals for both augmenting and evaluating machine learning based NLP models (Artemova et al., 2020). The approaches taken in those studies can be categorized as encoding or decoding cognitive processing signals. Encoding and decoding are complementary operations: encoding uses stimuli to predict brain activity, while decoding uses the brain activity to predict information about the stimuli (Naselaris et al., 2011). In this chapter, we focus on the decoding process for predicting information about the text input.

Until now, mostly eye tracking and functional magnetic resonance imaging (fMRI) signals have been leveraged to this purpose (e.g., Fyshe et al. (2014)). Eye tracking is cheaper to record and is widely used in psycholinguistic studies; fMRI recordings are expensive, but provide high spatial resolution, which furthers the research of localization of language-related cognitive processes. In this work, we take a different path by analysing the potentials and benefits of decoding electroencephalography signals for NLP. Through decoding EEG signals, we aim to explore the specific mental tasks occurring during language understanding, more specifically, during English sentence comprehension. More than a practical application of improving real-world NLP tasks, our main goal is to explore to what extent there is additional linguistic processing information in the EEG signal to complement the text input.

EEG is a non-invasive method to measure electrical brain activity. The synchronized activity of neurons in the brain produces electrical currents. The resulting voltage fluctuations can be recorded with external electrodes on the scalp. Compared to fMRI and other neuroimaging techniques, EEG can be recorded with a very high temporal resolution. This precision enables more naturalistic language understanding experiments, which is crucial for applications in NLP (Alday, 2019; Pfeiffer et al., 2020). Additionally, EEG is a less expensive and more easily available recording technique than fMRI. Compared to eye tracking, EEG may be more cumbersome to record and requires more expertise, however, while eye movements indirectly reflect the cognitive load of text processing, EEG contains more direct and comprehensive brain activity involved in language processing.

Due to the complexity and the low signal-to-noise ratio in the EEG data, it is very challenging to isolate specific cognitive processes, more and more researchers are relying on machine learning techniques to decode the EEG signals (Craik et al., 2019). In this thesis, we are the first to investigate the potential of leveraging EEG signals for improving NLP tasks. We presented a first approach in the previous chapter, however, no systematic large-scale study for using EEG features for NLP has been conducted yet.

With the purpose of making language decoding studies from brain activity more interpretable, we follow the recommendations of Gauthier and Ivanova (2018): (1) We commit to a specific mechanism and task, and (2) subdivide the input feature space including theoretically founded preprocessing steps. We investigate the impact of augmenting recurrent neural networks for sequence classification tasks with a range of EEG features, with a varying degree of theory-driven and data-driven feature extraction. To analyze the impact of different EEG features, we perform experiments on multiple semantic language understanding tasks, namely sentiment analysis as a binary or ternary sentence classification, and relation detection as a multi-class and multi-label classification task.

We systematically analyze the potential of decoding electrical brain activity data for improving NLP tasks. We analyze the effect of augmenting NLP models with brain activity data in a large-scale study accounting for various dimensions:

1. We present a comparison of full EEG signals to more theory-driven feature extraction by splitting the word-level EEG features into *frequency bands*.
2. We contrast the effects of EEG features on multiple text *embedding types*: randomly initialized, GloVe and BERT embeddings.
3. We compare the improvements of EEG features on various *training data sizes*.
4. We analyze the impact of EEG features on varying *classification complexity*: from binary classification to multi-class and multi-label tasks.

This comprehensive study is completed by comparing the impact of the decoded EEG signals not only to a *text-only* baseline, but also baselines augmented with eye tracking as well as random noise. The results show substantial improvements on both sentiment analysis tasks, however no improvements achieved on relation detection. EEG performs better than, or at least on pair with eye tracking in many scenarios. This study shows the potential of decoding EEG for NLP, and provides a good basis for future studies.

Electroencephalography in NLP

Recordings of brain activity play an important role in furthering our understanding of how human language works (Murphy et al., 2018; Ling et al., 2019). The appeal and added value of using brain activity signals in linguistic research are intelligible (Stemmer and Connolly, 2012). Moreover, more datasets of cognitive processing signals in naturalistic experiment paradigms with real-word language understanding tasks are becoming available (Kandylaki and Bornkessel-Schlesewsky, 2019). Nevertheless, as of yet, the related work leveraging EEG signals for NLP is very limited and will be summarized below. Deep learning has been used often to decode EEG signals (see Craik et al. (2019) for a review), mostly for brain-computer interface technologies (e.g., Nurse et al. (2016)). In other domains, frequency band features are also widely used to narrow the EEG signals. However, this avenue has not yet been explored when leveraging EEG signals to enhance NLP models.

Eye tracking and EEG are complementary measures of cognitive load. Reading times on words in a sentence depend on the amount of information the words convey. This correlation can be observed in eye tracking data, but also in EEG data (Frank et al., 2015). Multiple studies of co-registration (i.e., recording two modalities simultaneously, such as EEG and eye tracking in the ZuCo corpus) have shown advantages for studying language comprehension under naturalistic conditions (e.g., Dimigen et al. (2011) and Henderson et al. (2013). Moreover, word predictability and semantic similarity show distinct patterns of brain activity during language comprehension (Frank and Willems, 2017). Murphy and Poesio (2010) showed that semantic categories can be detected in simultaneous EEG recordings. Muttenthaler et al. (2020) use EEG signals to train an attention mechanism, similar to Barrett et al. (2018a), who leveraged eye tracking signals to induce machine attention with human attention. However, EEG has not yet been leveraged for higher-level semantic tasks such as sentiment analysis and relation detection.

Feature Extraction When using eye tracking features in NLP, it is common to extract well-established theory-driven features (Hollenstein et al., 2020a; Mathias et al., 2020). These established metrics are derived from a large body of psycholinguistic research. However, the amount of cognitive processes and noise included in brain activity signals, make feature engineering much harder on fMRI and EEG signals. Magnetoencephalography and fMRI have recently been used to evaluate language models. For instance, Schwartz et al. (2019) fine-tune a BERT language model with brain activity which yields better predictions of brain activity and does not

harm the model's performance on downstream NLP tasks. Machine learning studies leveraging fMRI data also rely on standard preprocessing steps such as motion correction and spatial smoothing and then use data-driven approaches to reduce the number of features, e.g., principal component analysis. However, fMRI data is most often used over full sentences or longer phrases, since the extraction of word-level signals is more complex than for EEG due to the lower temporal resolution and hemodynamic delay.

4.2 Multi-Modal Machine Learning Framework

Linzen (2020) advocates for the grounding of NLP models in multi-modal settings to compare the generalization abilities of the model to human language learning. Developing models that learn from such multi-modal input efficiently is crucial to advance the generalization abilities of state-of-the-art NLP models. Leveraging EEG and other cognitive processing data seems especially appealing to model multi-modal human-like learning processes. Previous work using cognitive data for improving NLP tasks mostly implement early fusion multi-modal methods, i.e., directly concatenating the textual and cognitive embeddings before inputting them into the network, e.g., our own experiments in Chapter 3, Barrett et al. (2018b) or Mishra et al. (2017a); or use the cognitive features to train an attention mechanism (Barrett et al., 2018a; Long et al., 2017). However, recent multi-modal machine learning work has shown the benefits of late fusion (Ramachandram and Taylor, 2017). Hence, we adopted this strategy in our work. In this section, we describe our proposed multi-modal machine learning framework, which learns simultaneously from text and from cognitive data such as eye tracking and EEG signals.

4.2.1 Data

With the purpose of augmenting information extraction tasks with brain activity signals, we leverage the two ZuCo datasets of simultaneous eye tracking and EEG data tailored to NLP problems, which we presented in Chapter 2. We select the normal reading paradigms from both corpora, in which participants were instructed to read English sentences in their own pace with no specific task beyond reading comprehension. The participants read one sentence at a time, using a control pad to move to the next sentence. This setup facilitated the natural reading

	ZuCo 1.0 Task SR	**ZuCo 1.0** Task NR	**ZuCo 2.0** Task NR
Participants	12	12	18
Sentences	400	300	344
Sentiment Analysis	✓	-	-
Relation Detection	-	✓	✓

Table 4.1: Details about the ZuCo tasks used in this chapter. In *Task SR* participants read sentences from movie reviews, and in *Task NR* sentences from Wikipedia articles.

paradigm. Details about the dataset sizes used to train and test the models for all the tasks are presented in Table 4.1. We split the data into sets of 80% for training and 20% for testing.

EEG Features

High-density EEG data were recorded using a 128-channel EEG device. Standard preprocessing steps for EEG data were applied, including band-pass filtering, artifact removal (i.e., removing blinks and other muscle activity), and quality assessment. As described in Chapter 2 ZuCo 1.0 and ZuCo 2.0 were preprocessed identically. During this phase, 23 electrodes in the outermost circumference (chin and neck) were used to detect muscular artifacts and were removed for subsequent analyses. Thus, each EEG data point, corresponding to a timestep of 2 milliseconds at a sampling rate of 500 Hz, contains 105 electrode values.

The fact that ZuCo provides simultaneous EEG and eye tracking data highly facilitates the extraction of word-level signals. Dimigen et al. (2011) demonstrated that EEG indices of semantic processing can be obtained in natural reading and compared to eye movement behavior. The eye tracking data provides millisecond-accurate fixation times for each word, so we obtain the brain activity during all fixations of a word.

In this chapter, we define the *full EEG signal* as the averaged brain activity over all fixations of a word, i.e., its total reading time. As described above, standard preprocessing steps have been applied to these signals. The EEG features consist of vectors of 105 dimensions, one for each electrode. We compare the full EEG features, a data-driven feature extraction approach, to *frequency band features*, a more theory-driven approach. We split the EEG signal into four frequency bands to limit the bandwidth of the EEG signals to be analyzed. The frequency bands

are fixed ranges of wave frequencies and amplitudes over a time scale: *theta* (4-8 Hz), *alpha* (8.5-13 Hz), *beta* (13.5-30 Hz), and *gamma* (30.5-49.5 Hz).

These frequency ranges are known to be associated with certain cognitive functions. Different neurocognitive aspects of language processing are associated with brain oscillations at various frequencies. *theta* activity reflects cognitive control and working memory (Williams et al., 2019), and increases when processing semantic anomalies (Prystauka and Lewis, 2019). Moreover, Bastiaansen et al. (2002) showed a frequency-specific increase in theta power as a sentence unfolds, possibly related to the formation of an episodic memory trace, or to incremental verbal working memory load. *Alpha* activity has been related to attentiveness (Klimesch, 2012). Both theta and alpha ranges are sensitive to the lexical–semantic processes involved in language translation (Grabner et al., 2007). *Beta* activity has been involved in higher-order linguistic functions such as the discrimination of word categories and the retrieval of action semantics as well as semantic memory, and syntactic processes, which support meaning construction during sentence processing. There is evidence that suggests that beta frequencies are important for linking past and present input and the detection of novelty of stimuli, which are essential processes for language perception as well as production (Weiss and Mueller, 2012). Beta frequencies also affect decisions regarding relevance (Eugster et al., 2014). Emotional processing of pictures enhances *gamma* band power (Müller et al., 1999). Gamma-band activity has been used to detect emotions (Li and Lu, 2009), and increases during syntactic and semantic structure building (Prystauka and Lewis, 2019). In the gamma frequency band, a power increase was observed during the processing of correct sentences, but this effect was absent following semantic violations (Hald et al., 2006). Frequency band features have often been used in deep learning methods for decoding EEG in other domains, such as mental workload and sleep stage classification (Craik et al., 2019).

As we showed in Chapter 3, and previous researchers have also shown (Foster et al., 2018), averaging over the EEG features of all subjects yield results almost as good as the single best-performing subjects. Hence, analogously to the eye tracking features, we also average the EEG features over all subjects to obtain more robust features.

Eye tracking Features

As we showed in the previous chapter, the cognitive language processing signals found in eye tracking data can be used to achieve modest, yet consistent improvements across a range of NLP tasks, including sentiment analysis and relation extraction. Since the ZuCo datasets provide si-

multaneous EEG and eye tracking recordings, we leverage the gaze data to augment all NLP tasks with eye tracking features as an additional multi-modal cognitive baseline. The gaze features used in this chapter are the same as provided in the ZuCo corpus (Chapter 2): number of fixations, first fixation duration, total reading time, gaze duration and go-past time.

4.2.2 Tasks

In this section, we describe the three sequence classification tasks for information extraction, on which we test the multi-modal EEG models.

Sentiment Analysis Analogous to the previous chapter, we compare binary (*positive/negative*) and ternary (+ *neutral*) sentiment classification on sentence level. For this task, we leverage only the sentences recorded in the first task of ZuCo 1.0, since they are part of the Stanford Sentiment Treebank (Socher et al., 2013), and thus directly provide annotated sentiment labels. All the 400 sentences are used for the ternary classification task, whereas neutral sentences are dropped for the binary classification resulting in 263 sentences.

Relation Detection The ZuCo corpus also contains Wikipedia sentences with relation types such as *Job Title*, *Nationality* and *Political Affiliation*. The sentences in ZuCo 1.0 and ZuCo 2.0, for the normal reading experiment paradigms, were initially extracted from the Wikipedia dataset provided by Culotta et al. (2006) and include 11 of the original relation types. In order to increase the task complexity, we treat this task differently than in the previous chapter. Since any sentence can include zero, one or more of the relevant semantic relations, we treat relation detection as a multi-class *and* multi-label sequence classification task. Removing duplicates between ZuCo 1.0 and ZuCo 2.0 resulted in 594 sentences used for training the models. Figure 4.1 illustrates the label and relation distribution of the sentences used to train the relation detection task[1].

[1]Note that we refer to this task as *relation detection*, while in Chapter 3 the *relation classification* task was defined as a single-label classification.

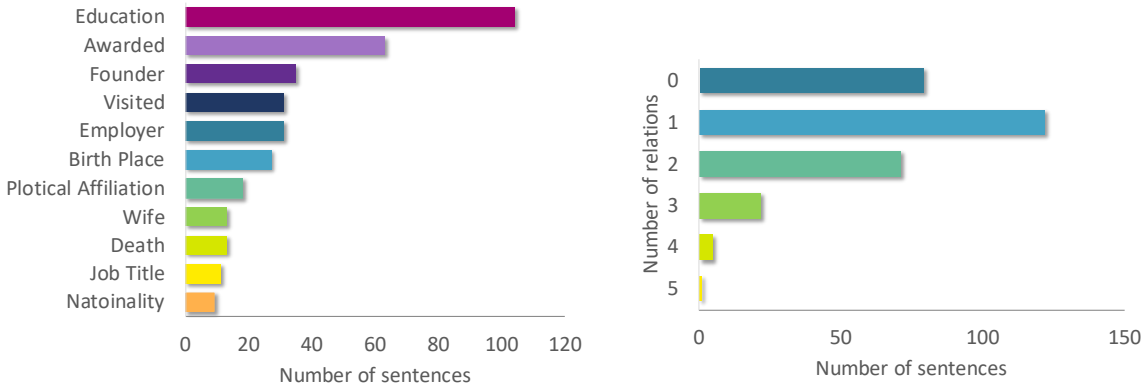


Figure 4.1: (left) Label distribution of relation detection dataset. (right) Distribution of relation types in the relation detection dataset.

4.2.3 Models

We present a multi-modal recurrent neural architecture to augment NLP sequence classification tasks with any other type of data. The goal of the multi-modal model is to find complementary information between different feature modalities (e.g., textual and EEG) in order to achieve a higher accuracy on the task at hand. Though combining different modalities or types of information for improving performance seems an intuitively appealing task, in practice, it is challenging to combine the varying level of noise and conflicts between modalities. We present multi-modal models for various NLP tasks, combining the learned representations of all input types (i.e., text and EEG features) in a late fusion mechanism before the final classification. Purposefully, this gives equal weight to all input modalities, which is sensible since we want to be able to assess the impact of the EEG data.

Such multi-modal models have been successfully applied in other domains, mostly across domains, for instance, learning speech reconstruction from silent videos (Ephrat et al., 2017), or for text classification using images (Kiela et al., 2018). Do et al. (2017) argument in favor of concatenating the hidden layers instead of concatenating the features at input time. In their review Ramachandram and Taylor (2017) show that many studies on multi-modal learning chose late fusion. Following, we describe the unimodal and multi-modal baseline models, as well as the multi-modal NLP models that jointly learn from text and human brain activity.

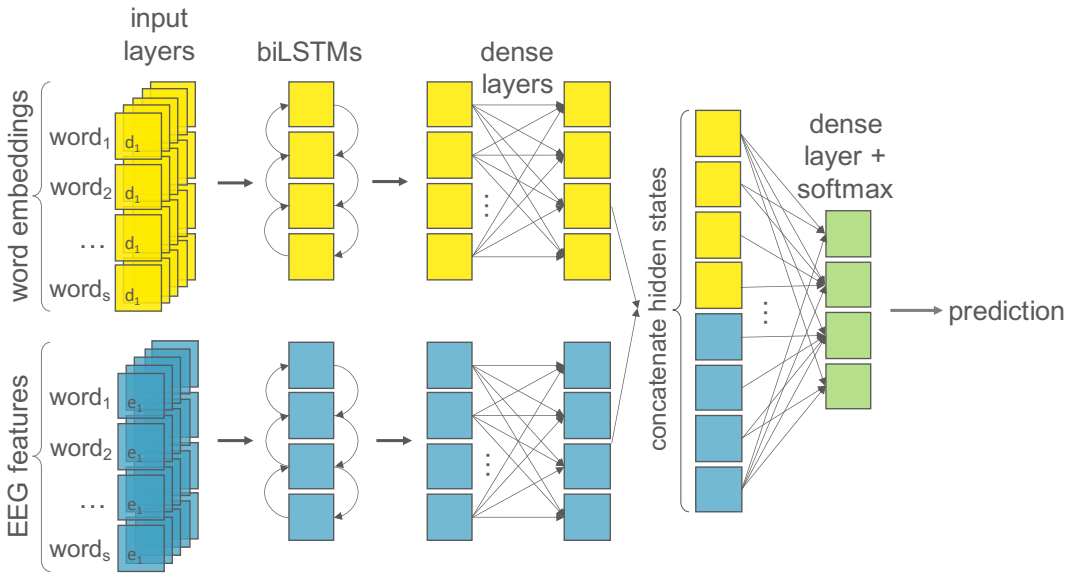


Figure 4.2: Multi-modal architecture for the EEG-augmented models. Word embeddings of dimension d are the input for the textual component (yellow); EEG features of dimension e for the cognitive component (blue). Both components consist of recurrent layers followed by two dense layers with dropout. Finally, the hidden states of both components are concatenated and followed by a final dense layer with softmax activation for classification (green).

Text Baselines

For each of the tasks presented above, we train unimodal models on word embedding features only. To analyze the interplay between various types of word embeddings and EEG data, we use the following three embedding types typically used in practice: randomly initialized embeddings (32 dimensions) trained at run time on the sentences provided, GloVe pre-trained embeddings of 300 dimensions based on word co-occurrence statistics (Pennington et al., 2014), and BERT pre-trained contextual embeddings (uncased base model of 768 dimensions with attention masks; Devlin et al. (2019)).

The randomly initialized word representations define word embeddings as n-by-d matrices, where n is the vocabulary size and d the embedding dimension, composed entirely of random values. Non-contextual word embeddings such as GloVe encode each word in a fixed vocabulary as a vector. The purpose of these vectors is to encode semantic information about a word, such that words with similar meanings result in similar embedding vectors. BERT consists of multiple layers of transformers with self-attention (Vaswani et al., 2017). Given a sentence,

BERT encodes each token into a feature vector which incorporates information from the token's context in the sentence.

The text baseline model consists of a first layer taking the embedding as an input, followed by a bidirectional LSTM, than two fully-connected dense layers with dropout between them, and finally a prediction layer using softmax activation. This corresponds to a single component of our multi-modal architecture (i.e., the top component in Figure 4.2). Following best practices (e.g., Sun et al. (2019)), we set BERT and the randomly initialized embeddings to be trainable, resulting in fine-tuning. However, the GloVe embeddings are fixed to the pre-trained weights and thus do not change during training.

Multi-modal Baselines

To analyze the effectiveness our multi-modal architecture with EEG signals properly, we not only compare to uni-modal text baselines, but also to multi-modal baselines using the same architecture described in the next section for the EEG models: (1) Gaze augmented baselines, where the five eye tracking features described in Section 3.2.1 are combined with the word embeddings by adding them to the multi-modal model in the same manner as the EEG features, as vectors with dimension = 5. (2) Random noise augmented baselines, where we add uniformly sampled vectors of random numbers as the second input data type to the multi-modal model. These random vectors are of the same dimension as the EEG vectors (i.e., dimension = 105).

EEG Models

To fully understand the impact of the EEG data on the NLP models, we build a model that is able to deal with multiple inputs and mixed data. We present a multi-modal model with late, decision-level fusion to learn joint representations of textual and cognitive input features. Figure 4.2 depicts the chosen architecture.

All input sentences are padded to the maximum sentence length to provide fixed-length text inputs to the model. Word embeddings of dimension d are the input for the textual component, where $d = 32, 300, 768$ for randomly initialize, Glove and BERT embeddings, respectively. EEG features of dimension e are the input for the cognitive component, where $e = 105$. Both components consist of bidirectional LSTM layers followed by two dense layers with dropout. Text and EEG features are given as independent inputs to their own respective component of the

network. The hidden representations of these are then concatenated before being fed to a final dense classification layer.[2] Although the goal of each networks is to learn feature transformations for their own modality, the relevant extracted information should be complementary. This is achieved, as commonly done in deep learning, through alternatively running inference and back-propagation of the data through the entire network enabling information to *flow* from the component responsible for one input modality to the other via the fully connected output layer. In order to learn a non-linear transformation function for each component, we employ the ReLU activation function after each hidden layer.

For binary and ternary sentiment analysis, the final dense layer has a softmax activation for classification. For the multi-label classification case of relation detection, we replace the softmax function in the last dense layer of the model with a sigmoid activation to produce independent scores for each class. If the score for any class surpasses a certain threshold, the sentence is labeled to contain that relation type (opposite to simply taking the *max* score as the label of the sentence). The threshold is tuned as an additional hyper-parameter.

This multi-modal model with separate components learned for each input data type has several advantages: It allows for separate pre-processing of each type of data, e.g., it can deal with differing tokenization strategies, which is useful in our case since it is challenging to map linguistic tokenization to the word boundaries presented to participants during the recordings of eye tracking and brain activity. Moreover, this approach is scalable to any number of input types. The generalizability of our model enables the integration of multiple data representations, e.g., learning from brain activity, eye movements, and other cognitive modalities simultaneously.

Hyper-parameters

To assess the impact of the EEG signals under fair modelling conditions, hyper-parameters are tuned for all baseline models as well as for all eye tracking and EEG augmented models. The ranges of the hyper-parameters are presented in Table 4.2. All results are reported as means over five runs with different random seeds. In each run, 5-fold cross validation is performed on a 80% - 20% training and test split. The best parameters were selected according to the model's

[2]We also experimented with different merging mechanisms to join the text and EEG layers of our two-tower model (concatenation, addition, subtraction, maximum). Concatenation achieved the best results, so we report only these.

Parameter	Range
LSTM layer dimension	64,128,256,512
Number of LSTM layers	1, 2, 3, 4
Dense layer dimension	32, 64, 128, 256, 512
Dropout	0.1, 0.3, 0.5
Batch size	20, 40, 60
Learning rate	10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}
Threshold	0.3, 0.5, 0.7
Random seeds	13, 22, 42, 66, 78

Table 4.2: Tested ranges for all hyper-parameters. *Threshold* only applies to relation detection.

accuracy on the validation set (10% of the training set) across all 5 folds. We implemented early stopping with a patience of 80 epochs and a minimum delta in validation accuracy of 10^{-7}. The validation set is used for both parameter tuning and early stopping.

4.3 Results & Discussion

All results for the best set of hyper-parameters are presented in Table 4.3 for binary sentiment analysis, Table 4.4 for ternary sentiment analysis, and Table 4.5 for relation detection. As described above we select the best hyper-parameters based on the best validation accuracy achieved. For analyzing the results we focus on macro-averaged precision (P), recall (R) and F_1-score. The definition of the evaluation metrics can be found in Appendix A.1. In both sentiment tasks, the eye tracking and EEG data yields a modest but consistent improvement over the text baseline. However, in the case of relation detection the addition of either eye tracking or brain activity data seems to be harmful. Generally, the results show a decreasing maximal performance per task with increasing task complexity measured in terms of number of classes.

In this section, we discuss these results from different angles. We contrast the performance of different EEG features, we compare the EEG results to the text baselines and multi-modal baselines (as described in Section 4.2.3), and we analyze the effect of different embedding types. Additionally, we explore the impact of varying training set sizes in a data ablation study. Finally, we investigate the possible reasons for the decrease in performance for the relation detection task, which we associate with the task complexity.

Model	Randomly initialized			GloVe			BERT		
	P	**R**	**F_1 (std)**	**P**	**R**	**F_1 (std)**	**P**	**R**	**F_1 (std)**
Baseline	0.572	0.573	0.552 (0.07)	0.751	0.738	0.728 (0.08)	0.900	0.899	0.893 (0.04)
+ noise	0.599	0.574	0.541 (0.08)	0.721	0.715	0.709 (0.09)	0.921	0.918	0.915 (0.05)
+ ET	**0.615**	**0.605**	0.586 (0.06)	**0.795**	**0.786**	**0.781** (0.06)	0.913	0.907	0.904 (0.05)
+ EEG full	0.562	0.560	0.550 (0.07)	0.752	0.747	0.744 (0.04)	0.909	0.908	0.903 (0.05)
+ EEG θ	0.585	0.593	0.556 (0.10)	0.770	0.766	0.761 (0.07)	0.922	0.919	0.919 (0.03)
+ EEG α	0.609	0.601	**0.592** (0.06)	0.775	0.767	0.760 (0.06)	0.915	0.915	0.913 (0.04)
+ EEG β	0.610	0.589	0.563 (0.07)	0.781	0.773	0.770 (0.06)	0.914	0.914	0.911 (0.05)
+ EEG γ	0.565	0.580	0.559 (0.09)	0.775	0.766	0.763 (0.05)	**0.929**	**0.927**	**0.926** (0.03)
+θ+α+β+γ	0.568	0.563	0.528 (0.09)	0.783	0.779	0.776 (0.07)	0.923	0.925	0.92 (0.05)

Table 4.3: Precision (P), recall (R), F_1-score and the standard deviation (std) between runs for binary sentiment analysis. The best results per column are marked in bold, all EEG results better than the text baseline *and* the baseline augmented with random noise are marked with grey background.

Model	Randomly initialized			GloVe			BERT		
	P	**R**	**F_1 (std)**	**P**	**R**	**F_1 (std)**	**P**	**R**	**F_1 (std)**
Baseline	0.412	0.406	0.365 (0.08)	0.516	0.510	0.501 (0.04)	0.722	0.714	0.710 (0.05)
+ noise	0.373	0.399	0.344 (0.10)	0.531	0.519	0.504 (0.04)	0.711	0.706	0.700 (0.06)
+ ET	**0.424**	**0.413**	**0.388** (0.06)	0.539	**0.528**	0.513 (0.04)	0.728	0.717	0.714 (0.05)
+ EEG full	0.386	0.382	0.341 (0.05)	0.496	0.492	0.479 (0.08)	0.724	0.715	0.711 (0.06)
+ EEG θ	0.386	0.396	0.355 (0.09)	**0.543**	0.525	0.510 (0.07)	0.715	0.708	0.704 (0.05)
+ EEG α	0.409	0.410	0.379 (0.07)	0.509	0.507	0.491 (0.06)	0.720	0.712	0.707 (0.05)
+ EEG β	0.381	0.404	.359 (0.08)	0.535	0.523	0.512 (0.05)	**0.732**	**0.720**	**0.717** (0.07)
+ EEG γ	0.401	0.413	0.382 (0.06)	0.535	0.525	**0.515** (0.05)	0.709	0.705	0.697 (0.06)
+θ+α+β+γ	0.419	0.410	0.372 (0.08)	0.516	0.501	0.494 (0.08)	0.722	0.717	0.713 (0.05)

Table 4.4: Precision (P), recall (R), F_1-score and the standard deviation (std) between runs for ternary sentiment analysis. The best results per column are marked in bold, all EEG results better than the text baseline *and* the baseline augmented with random noise are marked with grey background.

EEG Feature Analysis

We start by investigating the impact of the different EEG features: full EEG signals (data-driven feature extraction), EEG signals filtered by frequency bands on word-level (minimally theory-

Model	Randomly initialized			GloVe			BERT		
	P	**R**	**F_1 (std)**	**P**	**R**	**F_1 (std)**	**P**	**R**	**F_1 (std)**
Baseline	**0.569**	**0.382**	**0.453** (0.04)	**0.658**	0.511	**0.573** (0.04)	**0.733**	**0.732**	**0.731** (0.03)
+ noise	0.462	0.335	0.382 (0.05)	0.577	0.497	0.532 (0.03)	0.675	0.585	0.625 (0.03)
+ ET	0.468	0.324	0.373 (0.06)	0.547	0.476	0.506 (0.04)	0.661	0.631	0.644 (0.03)
+ EEG full	0.426	0.335	0.370 (0.06)	0.519	0.449	0.480 (0.05)	0.677	0.627	0.650 (0.03)
+ EEG θ	0.474	0.354	0.402 (0.05)	0.574	0.503	0.534 (0.04)	0.700	0.650	0.673 (0.03)
+ EEG α	0.497	0.338	0.392 (0.05)	0.580	**0.521**	0.548 (0.03)	0.694	0.652	0.671 (0.03)
+ EEG β	0.446	0.360	0.394 (0.05)	0.565	0.475	0.507 (0.06)	0.702	0.656	0.677 (0.03)
+ EEG γ	0.448	0.347	0.386 (0.04)	0.562	0.497	0.526 (0.04)	0.690	0.652	0.669 (0.03)
+ θ+α+β+γ	0.472	0.320	0.370 (0.06)	0.470	0.399	0.426 (0.06)	0.675	0.646	0.659 (0.04)

Table 4.5: Precision (P), recall (R), F_1-score and the standard deviation (std) between runs for relation detection. The best results per column are marked in bold, all EEG results better than the text baseline *and* the baseline augmented with random noise are marked with grey background.

driven feature extraction), and EEG signals filtered by frequency bands only during the first fixation of a word (extended theory-driven feature extraction).

Results show that our multi-modal models yield better results with filtered EEG frequency bands than using the full EEG signal on all tasks. Although the alpha and gamma features show promising results on some embedding types and tasks (e.g., BERT embeddings and gamma features for binary sentiment analysis reported in Table 4.3), the results show no clear sign of any frequency band outperforming the others (neither across tasks for a fixed embedding type, nor for a fixed task and across all embedding types). Moreover, the combination of all four EEG frequency bands (i.e., in a multi-modal model of 5 components, including text embeddings), does not further increase the results, except for binary sentiment analysis with GloVe embeddings. Hence, further exploring the effects of specific frequency bands on language understanding tasks might prove useful.

In addition to the full EEG and frequency band features on word-level, we test a more theory-driven approach for feature extraction for both sentiment analysis tasks. Investigations into the neural dynamics of reading sentences conveying a sentiment have shed light on the cognitive processing of positive and negative words. Pfeiffer et al. (2020) found sentiment-related differences in brain activity at 224–304 ms after fixation onset. Therefore, we approximate this finding and restrict the EEG features to those occurring within the duration of the first fixation

Model	TRT	FFD	Difference (%)
Binary sentiment			
Random init. + γ	0.382	0.358	-2.37
GloVe + γ	0.515	0.496	-1.91
BERT + β	0.717	0.717	-0.04
Ternary sentiment			
Random init. + α	0.592	0.579	-1.31
GloVe + β	0.770	0.757	-1.37
BERT + γ	0.926	0.916	-0.98

Table 4.6: F_1-scores for the best sentiment analysis model combinations with EEG features during the total reading time (TRT) of a word, and EEG features covering only the first fixation duration (FFD) of a word.

(ranging between 100 and 700 ms), where the sentiment processing occurs. Then, we trained the EEG models for the best embedding + frequency band combinations on these theory-driven features. However, the results were inferior to using the EEG features during the total reading time of a word (see Table 4.6).

Focusing on first-fixation EEG features might align better with the theory, but hinders the deep model to extract as much useful information as when filtering the EEG features based on the total reading time of a word. Data-driven methods can help us to tease more information from the recordings by allowing to test broader theories and task-specific language representations (Murphy et al., 2018).

Comparison to Multi-Modal Baselines

The multi-modal EEG models often outperform the text baselines (at least for the sentiment analysis tasks). We now analyze how the EEG models compare to the two augmented baselines described in Section 4.2.3 (i.e., eye tracking and models augmented with random noise). We find that EEG always performs better or equal to the multi-modal text + eye tracking models. This shows how promising EEG is as a data source for multi-modal cognitive NLP. Although eye tracking requires less recording efforts, these results corroborate that EEG data contain more information about the cognitive processes occurring in the brain during language understanding.

As expected, the baselines augmented with random noise perform worse than the pure text baselines in all cases expect for binary sentiment analysis with BERT embeddings. This model seems to deal exceptionally well with added noise. In the case of relation detection, the added noise harms the models similarly to adding EEG. It becomes clear for this task that adding the full EEG features is worse than adding random noise, but some of the frequency band features clearly outperform the augmented noise baseline.

Comparison of Embedding Types

Our baseline results show that contextual embeddings outperform the non-contextual methods across all the tasks. Arora et al. (2020) also compare random, GloVe and BERT embeddings and find that with smaller training sets, the difference in performance between these three embedding types is larger. This is in accordance with our results, which show that the type of embedding has a large impact on the baseline performance on all three tasks. The improvements of EEG in the binary sentiment analysis task with BERT embeddings are especially noteworthy.

Augmenting our baseline with EEG data on the binary sentiment analysis tasks results in approximately +3% F1 score across all the different embeddings. The gain is slightly lower at +1% for all the embeddings in the ternary sentiment classification task. While there is no gain for relation detection, the differences is also constant across embeddings. This shows that the improvements gained by adding EEG signals are much more dependent on the task than on the embedding type. In foresight, this finding might be useful in the future, when new embeddings will improve the baseline performance even further while upholding the stable gain from the EEG signals.

Data Ablation

One of the challenges of NLP is to learn as much as possible from limited resources. Unlike most machine learning models, one of the most striking aspects of human learning is the ability to learn new words or concepts from limited numbers of examples (Lake et al., 2015). Using cognitive language processing data may allow us take a step towards meta-learning, the process of discovering the cognitive processes that are used to tackle a task in the human brain (Griffiths et al., 2019), and in turn be able to improve the generalization abilities of NLP models. Humans can learn from very few examples, while machines, particularly deep learning models, typically

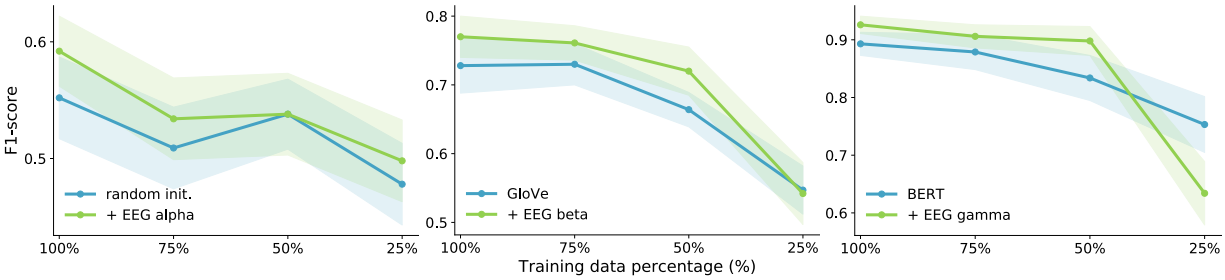


Figure 4.3: Data ablation for all three embedding types for binary sentiment analysis. The shaded areas represent the standard deviations.

need many examples. Perhaps this advantage of humans is due to their multi-modal learning mechanisms (Linzen, 2020).

Therefore, we analyze the impact of adding EEG features to our NLP models with less training data. We performed data ablation experiments for all three tasks. The most conclusive results were achieved on binary sentiment analysis. Randomly initialised embeddings unsurprisingly suffer a lot when reducing training data. The results are shown in Figure 4.3. We present the results for the best-performing frequency bands only. GloVe and BERT embeddings yield the largest gain from EEG data is obtained with only 50% of the training data which is as little as 105 training sentences. These experiments emphasize the potential of EEG signals for NLP especially when dealing with very small amounts of training data.

Task Complexity Ablation

From the previously described results, one hypothesis on the reason why augmenting the baseline with either EEG or eye tracking data lowers the performance in the relation detection task lies in the complexity of the task. More concretely, we measure the complexity by counting the number of classes the model needs to learn. We see a decreasing performance boost with increasing complexity over the three evaluated tasks. Therefore, we validate this hypothesis by simplifying the relation detection task by reducing the number of classes from 11 to 2. We create three binary relation detection tasks for the two most frequent relation types, *Job Title* and *Visited* (see Figure 4.1. For example, we classify all the samples containing the relation *Job Title* (184 samples) against all samples with no relation (219 samples).

We train these additional models with BERT embeddings. The results for full EEG features and the best frequency band from the previous results are shown in Table 4.7. It is evident that with the simplification of the relation detection task into binary classification tasks, EEG signals

Job Title	Precision	Recall	F_1-score
BERT	0.870	0.871	0.868 (0.03)
BERT + EEG full	**0.899**	**0.900**	**0.897** (0.03)
BERT + EEG β	0.897	0.899	**0.897** (0.03)
Visited	**Precision**	**Recall**	**F_1-score**
BERT	0.864	0.848	0.848 (0.05)
BERT + EEG full	**0.884**	0.863	0.864 (0.04)
BERT + EEG β	0.883	**0.869**	**0.866** (0.06)

Table 4.7: Binary relation detection results. The best result in each column is marked in bold.

boost the performance and achieve considerable improvements over the text baseline. The gains are similar as for binary sentiment analysis. This confirms our hypothesis that the EEG features tested yield good results on simple tasks, but more research is needed to achieve improvements on more complex tasks.

4.4 Summary

We presented a large-scale study about leveraging electrical brain activity signals during reading comprehension for augmenting machine learning models of semantic language understanding tasks, namely sentiment analysis and relation detection. We analyzed the effects of different EEG features and compared the multi-modal models to multiple baselines. Moreover, we compared the improvements gained from the EEG signals on three different types of word embeddings. Not only did we test the effect of varying training set sizes, but also tasks of various difficulty levels (in terms of number of classes).

We achieve consistent improvements with EEG across all three embedding types, but the improvement magnitude decreases for more difficult tasks. While the improvement for the binary and ternary sentiment analysis tasks range between 1-4% F-score, for relation detection, a multi-class and multi-label sequence classification task, it was not possible to achieve any improvements unless the task complexity is substantially reduced.

To sum up, we capitalize on the advantages of electroencephalography data to examine if and which EEG features can serve to augment language understanding models. While our results

show that there is linguistic information in the EEG signal complementing the text features, more research is needed to isolate language-specific features. More generally, the work in this thesis paves the way for in-depth EEG-based NLP studies.

Chapter 5

Predicting Human Reading Behavior

In this chapter, we investigate to what extent human reading behavior can be predicted by state-of-the-art pre-trained language models. It has been established that combining insights from neuroscience and machine learning will take us closer to human-level language understanding (McClelland et al., 2020). As we have shown in Chapter 3, eye movement data has been used in NLP to provide inductive biases to models to achieve improvements on certain tasks. It has also been used as a supervisory signal in neural networks for NLP in attention mechanisms (Barrett et al., 2018a) and in multi-task learning scenarios (Klerke et al., 2016; Gonzalez-Garduno and Søgaard, 2017).

Therefore, we compare the performance of language-specific and multilingual pre-trained transformer models to predict reading time measures reflecting natural human sentence processing on Dutch, English, German, and Russian texts. This results in accurate models of human reading behavior and yields insights into the workings of transformer language models. First, we describe the relevant prior work in modelling human reading behavior and give some background about probing transformer language models in Section 5.1. In Section 5.2, we describe the eye tracking data and language models used in this work and we outline our method. Finally, in a series of experiments presented in Section 5.3, we analyze the cross-domain and cross-language capabilities of these models and show how they reflect human sentence processing.

The contents of this chapter are based on the following publication: Hollenstein, N., Pirovano, F., Zhang, C., Jäger, L., & Beinborn, L. Multilingual language models predict human reading behavior. *Under review*.

5.1 Background

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

In human language processing, phenomena of relative importance can be approximated indirectly by tracking eye movements and measuring fixation durations (Rayner, 1977). It has been shown that fixation durations and relative importance of text segments are strongly correlated in natural reading, so that direct links can be established on the token level (Malmaud et al., 2020). In the example in Figure 5.1, the newly introduced entity *Mary French* is fixated twice and for a longer duration because it is relatively more important for the reader than the entity *Laurence* which had been introduced in the previous sentence. Being able to reliably predict eye movement patterns from the language input would bring us one step closer to unraveling the underlying cognitive processes of language understanding.

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

While psycholinguistic work mainly focuses on very specific phenomena of human language processing that are typically tested in experimental settings with constructed stimuli (Hale, 2017), we focus on directly generating token-level predictions from natural reading. As we showed in Chapter 3, the performance of computational models can be improved even further if their inductive bias is adjusted using human cognitive signals such as eye tracking, fMRI, or

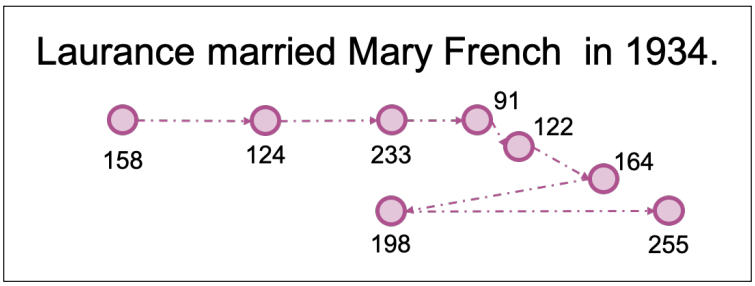


Figure 5.1: From the fixation times in milliseconds of a single subject in the ZuCo 1.0 dataset, the feature vectors described in Section 5.2.1 for the words "Mary" and "French" would be, respectively: $[2, 233, 233, 431, 215.5, 1, 1, 1]$ and $[2, 91, 213, 213, 106.5, 1, 1, 1]$.

EEG data (Toneva and Wehbe, 2019; Takmaz et al., 2020).

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

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

Probing Transformer Language Models

Contextualized neural language models have become increasingly popular, but our understanding of these black box algorithms is still rather limited (Gilpin et al., 2018). Current intrinsic

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

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

Modelling Human Sentence Processing

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

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

Compared to neural models of human sentence processing, we predict not only individual metrics, but a range of eye tracking features covering the full reading process from early lexical access to late syntactic processing. Most models of reading focus on predicting skipping probability (Reichle et al., 1998; Matthies and Søgaard, 2013; Hahn and Keller, 2016). These models achieve between 55 and 70% accuracy when predicting fixation probability. Sood et al. (2020b) propose a salience model to compute the attention scores for task-specific NLP models. Moreover, the results by Sood et al. (2020a) suggest that different architectures seem to learn rather different neural attention strategies and similarity of neural to human attention does not guarantee best performance, which is in line with the findings from Jain and Wallace (2019).

5.2 Fine-Tuning Language Models on Eye Tracking Data

In this section, we describe the eye tracking corpora used for our experiments and the feature extraction. We then present the language models we use for fine-tuning. Lastly, we describe the training and evaluation procedure.

5.2.1 Data

We predict eye tracking data only from naturalistic reading studies, in which the participants read full sentences or longer spans of naturally occurring text at their own speed. The data from these studies exhibit higher ecological validity than neurolinguistic studies which rely on artificially constructed sentences and paced presentation (Alday, 2019). In order to conduct a cross-lingual comparison, we use eye tracking data collected from native speakers of four languages (see Table 5.1 for details):

1. **English:** The largest number of eye tracking data sources are available available for English. We leverage 3 English corpora providing gaze features during reading: (1) The Dundee corpus (Kennedy et al., 2003) contains 20 newspaper articles from *The Independent*, which were presented to English native readers on a screen five lines at a time. (2) The GECO corpus (Cop et al., 2017) contains eye tracking data from English monolinguals reading the entire novel *The Mysterious Affair at Styles* by Agatha Christie. The text

Corpus	Lang.	Subjs.	Sents.	Sent. length	Tokens	Types	Word length	Flesch
Dundee	EN	10	2,379	21.7 (1–87)	51,497	9,488	4.9 (1–20)	53.3
GECO	EN	14	5,373	10.5 (1–69)	56,410	5,916	4.6 (1–33)	77.4
ZuCo	EN	30	1,053	19.5 (1–68)	20,545	5,560	5.0 (1–29)	50.6
GECO	NL	19	5,190	11.64 (1–60)	59,716	5,575	4.5 (1–22)	57.5
PTC	DE	76	97	19.5 (5–51)	1,895	847	6.5 (2–33)	36.4
RSC	RU	103	144	9.4 (5–13)	1,357	993	5.7 (1–18)	64.7

Table 5.1: Descriptive statistics of all eye tracking datasets[3]. The last column shows the Flesch Reading Ease score Flesch (1948) which ranges from 0 to 100 (higher score indicates easier to read). Adaptations of the Flesch score were used for Dutch (NL), German (DE) and Russian (RU) (see Appendix A.3).

was presented on the screen in paragraphs. (3) The ZuCo corpus (see Chapter 2) includes eye tracking data of full sentences from movie reviews and Wikipedia articles.[1]

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

3. **German:** The Potsdam Textbook Corpus (PTC, (Makowski et al., 2018)) contains 12 short passages from college-level biology and physics textbooks, which are read by expert and laymen German native speakers. The full passages were presented on multiple lines on the screen.

4. **Russian:** The Russian Sentence Corpus (RSC, (Laurinavichyute et al., 2019)) contains 144 naturally occurring sentences extracted from the Russian National Corpus.[2] Full sentences were presented on the screen to monolingual Russian-speaking adults one at a time.

[1] We use the normal reading (NR) and sentiment reading (SR) tasks from ZuCo 1.0 and the normal reading (NR) task from ZuCo 2.0.

[2] `https://ruscorpora.ru`

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

Eye Tracking Features

Given a word *w* in a sentence, a fixation on *w* is defined as a period of time in which the eyes of the reader do not move from a region belonging to *w* (see Figure 5.1). Similar to our approach in Chapter 3, we predict the following eight eye tracking features for each token *w* in the input text. These encode the full reading process from early lexical access up to subsequent syntactic integration:

- **Word-level characteristics:** We extract basic features that encode *word-level* characteristics: (1) number of fixations (nFix) - the number of times a subject fixates *w*, averaged over all subjects; (2) mean fixation duration (MFD) - the average fixation duration of all fixations made on *w*, averaged over all subjects; (3) fixation proportion (FProp) - the number of subjects that fixated *w*, divided by the total number of subjects.
- **Early processing:** We include features to capture the *early* lexical and syntactic processing, based on the first time a word is fixated: (4) first fixation duration (FFD) - the duration, in milliseconds, of the first fixation on *w*, averaged over all subjects; (5) first pass duration (FPD) - the sum of all fixations on *w* from the first time a subject fixates *w* to the first time the subject fixates another token, averaged over all subjects.
- **Late processing:** We use measures reflecting the *late* syntactic processing and general disambiguation, based on words which were fixated more than once: (6) total reading time (TRT) - the sum of the duration of all fixations made on *w*, averaged over all subjects; (7) number of re-fixations (nRefix) - the number of times *w* is fixated after the first fixation, i.e., the maximum between 0 and the nFix-1, averaged over all subjects; (8) re-read proportion (ReProp) - the number of subjects that fixated *w* more than once, divided by the total number of subjects.

The values of these eye tracking features vary over different ranges. FFD, for example, is measured in milliseconds, and the average values are around 200 ms, whereas ReProp is a proportional measure, and therefore assumes floating-point values between 0 and 1. We standardize all eye tracking features independently (range: 0–100), so that the loss can be calculated uniformly over all feature dimensions.

Eye movements depend on the stimulus and as such are language-specific, but there exist universal tendencies which remain stable across languages. For example, the average fixation duration

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

Table 5.2: Pre-trained transformer language models analyzed in this chapter.

in reading ranges from 220 to 250 ms independent of the language. Furthermore, word characteristics such as word length, frequency, and predictability affect fixation duration similarly across languages, but the effect size depends on the language and the script (Laurinavichyute et al., 2019; Bai et al., 2008). The word length effect can be observed across all four languages included in this work (see Figure A.12 in Appendix A.3).

Language Models

We compare the ability to predict eye tracking features in two models: BERT and XLM. Both models are trained on the transformer architecture and yield state-of-the-art results for a wide range of NLP tasks (Liang et al., 2020). Multilingual BERT simply concatenates the Wikipedia input from 104 languages and is optimized by performing masked token and next sentence prediction as in the monolingual model (Devlin et al., 2019) without any cross-lingual constraints. In contrast, XLM explicitly uses parallel sentences in two languages as input to facilitate cross-lingual transfer (Lample and Conneau, 2019). Both BERT and XLM use subword tokenization methods to build shared vocabulary spaces across languages.

We use the pre-trained checkpoints from the HuggingFace repository for monolingual and multilingual models.[4] The details are reported in Table 5.2.

[4] https://huggingface.co/transformers/pre-trained_models.html

5.2.2 Method

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

Training Details We split the data into 90% training data, 5% validation, and 5% test data. We initially tuned the hyper-parameters manually and set the following values for all models: We use an AdamW optimizer (Loshchilov and Hutter, 2018) with a learning rate of 0.00005 and a weight decay of 0.01. The batch size varies depending on the model dimensions (Appendix A.3). We employ a linear learning rate decay schedule over the total number of training steps. We clip all gradients exceeding the maximal value of 1. We train the models for 100 epochs, with early stopping after 7 epochs without improvements on the validation accuracy.

Evaluation Procedure As the features have been standardized to the range 0–100, the *mean absolute error* (MAE; see Appendix A.1 for definition) can be interpreted as a percentage error. For readability, we report the prediction accuracy as 100−MAE. The results are averaged over batches and over 5 runs with varying random seeds. For a single batch of sentences, the overall MAE is calculated by concatenating the words in each sentence and the feature dimensions for each word, and padding to the maximum sentence length. The per-feature MAE is calculated by concatenating the words in each sentence. For example, for a batch of B sentences, each composed of L words, and G gaze features per word. Then, the overall MAE is calculated over a vector of $B*L*G$ dimensions, whereas the per-feature MAE is calculated over a vector of $B*L$ dimensions.

Baselines To test the language models' abilities on predicting human reading behavior only from pre-training on textual input, we take the provided model checkpoints and use them to predict the eye tracking features without any fine-tuning. We then compare these results to the results of the models fine-tuned on eye movement data to quantify the improvements. Moreover, on the individual feature level, we use a mean baseline to compare the prediction accuracy of

the language models to a baseline performance, where always the mean values of the original eye tracking features is predicted.

5.3 Results & Discussion

In this section we present the results of our experiments. First, we describe the general results, aggregated over all eye tracking features, as well as for individual features. Then, we discuss the cross-domain and cross-language evaluations and finally we analyze whether the fine-tuned language models reflect some of the properties observed in human reading behavior.

Tables 5.3 and 5.4 show that all fine-tuned models predict the eye tracking features with more than 90% accuracy for English and Dutch. For English, the BERT models yield high performance with standard deviation below 0.15 on all three datasets. The results for the XLM models are slightly better on average, but show much higher standard deviations. Similar to the results presented by Lample and Conneau (2019), we find that more training data from multiple languages improves the prediction performance. For instance, the XLM-100 model achieves higher accuracy than the XLM-17 model in all cases. For the smaller datasets, PTC and RSC, the multilingual models clearly outperform the monolingual models. For English, the differences are minor. More training data results in higher prediction accuracy even when the eye tracking data comes from various languages and was recorded in different reading studies by different devices (ALL-LANGS, fine-tuning on the data of all six datasets together). However, merely adding more data from the same language (ALL-EN, fine-tuning on the data from Dundee, GECO, and ZuCo together) does not result in higher performance.

To analyze this further, we perform an ablation study with varying amounts of training data. The results are shown in Figure 5.2 for Dutch and English. The performance of the XLM models remains stable even with a very small percentage of training data. The performance of the BERT models, however, drops drastically when fine-tuning on less than 20% of the data. Notably, even fine-tuning the XLM models on 1% of the eye tracking data already shows a large increase compared to the standard pre-trained models without fine-tuning. Similar to Merkx and Frank (2020) and Hao et al. (2020) we find that the model architecture, along with the composition and size of the training corpus have a significant impact on the psycholinguistic modeling performance.

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

Table 5.3: Prediction accuracy of the fine-tuned language models aggregated over all eye tracking features for the English corpora, including the concatenated dataset. Standard deviation is reported in parentheses; the best results marked in bold.

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

Table 5.4: Prediction accuracy of the fine-tuned language models aggregated over all eye tracking features for the Dutch, German and Russian corpora, and for all four languages combined in a single dataset. Standard deviation is reported in parentheses; the best results marked in bold.

Figure 5.3 presents the differences across models in predicting the individual gaze features.[5] Across all datasets, first pass duration (FPD) and number of re-fixations (NREFIX) are the most accurately predicted features. Proportions (FPROP and REPROP) are harder to predict because these features are even more dependent on subject-specific characteristics. Nevertheless, when comparing the prediction accuracy of each gaze feature to a baseline which always predicts the mean value, the predicted features FPROP and REPROP achieve the largest improvements relative to

[5] Plots for the remaining datasets can be found in Appendix A.3.

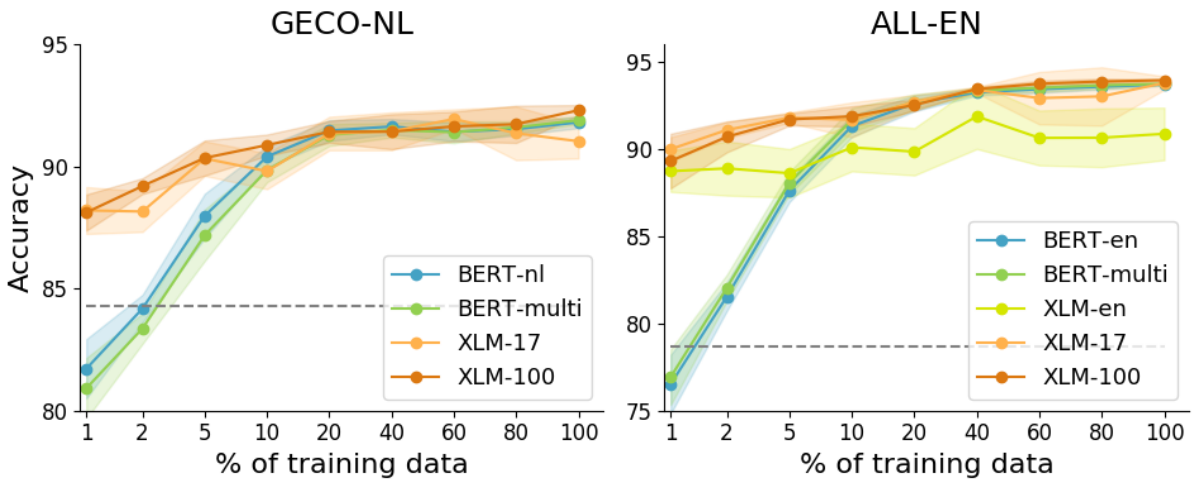


Figure 5.2: Data ablation study for Dutch, English and all languages together. The results are aggregated over all eye tracking features. In addition to the mean across five runs, the shaded areas represent the standard deviation. The dashed line is the result of the pre-trained BERT-Multi model without fine-tuning.

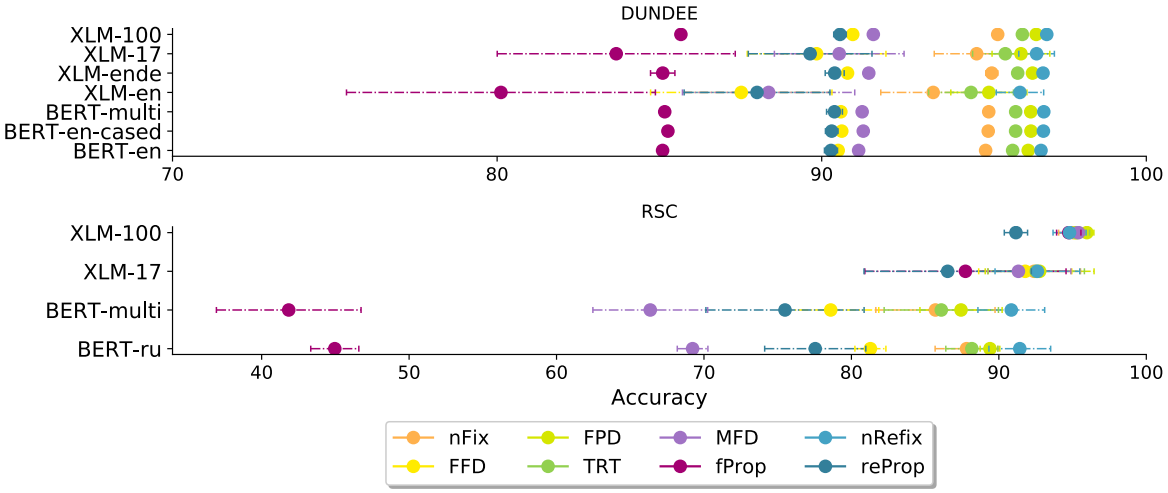


Figure 5.3: Results of individual eye tracking features (mean prediction accuracy and standard deviation bars) for all fine-tuned models on the Dundee (EN) and RSC (RU) corpora.

the mean baseline. See Figure 5.4 for a comparison between all features for the best performing model XLM-100 on all six datasets.

As described above, we also tested the performance of the pre-trained language models before fine-tuning. These results are presented in Tables A.3 and A.4 in Appendix A.3. The achieved

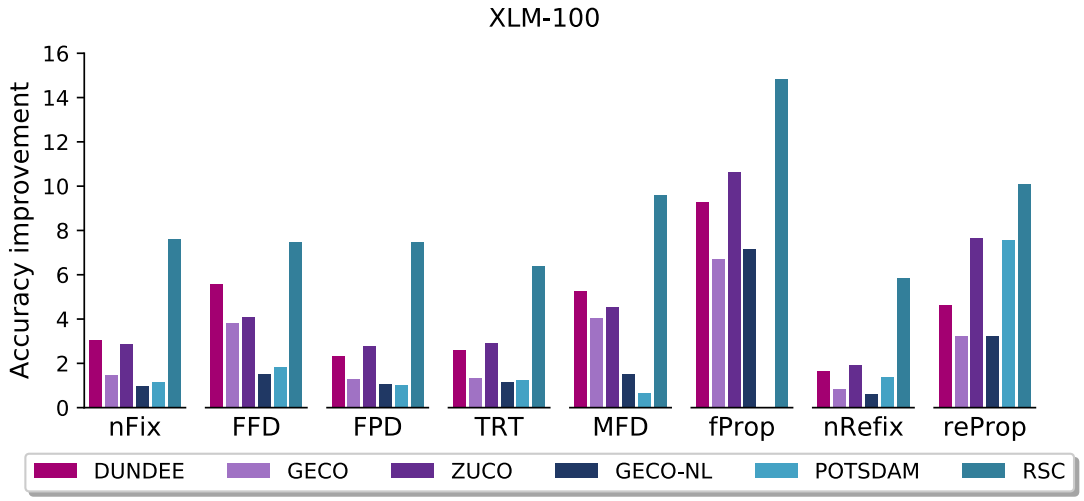


Figure 5.4: Improvement of prediction accuracy for the fine-tuned XLM-100 model predictions relative to the mean baseline for each eye tracking feature.

accuracy aggregated over all eye tracking features lies between 75-78% for English. For Dutch, the models achieve 84% accuracy, but for Russian merely 65%. Across the same languages the results between the different language models are only minimal. However, on the individual eye tracking features, the pre-trained models do not achieve any improvements over the mean baseline (Figure A.14).

Cross-Domain Evaluation

For the main experiment, we always tested the models on held-out data from the same dataset. Figure 5.5 shows the results of evaluating the eye tracking predictions on out-of-domain text for the English datasets. For instance, we fine-tune the model on the newspaper articles of the Dundee corpus and test on the literary novel of the GECO corpus. We see that the overall prediction accuracy across all eye tracking features is constantly above 90% in all combinations. This shows that our gaze prediction model is able to generalize across domains. We find that the cross-domain capabilities of BERT are slightly better than for XLM. BERT-En performs best in the cross-domain evaluation, possibly because its training data is more domain-general since it includes text from Wikipedia and books.

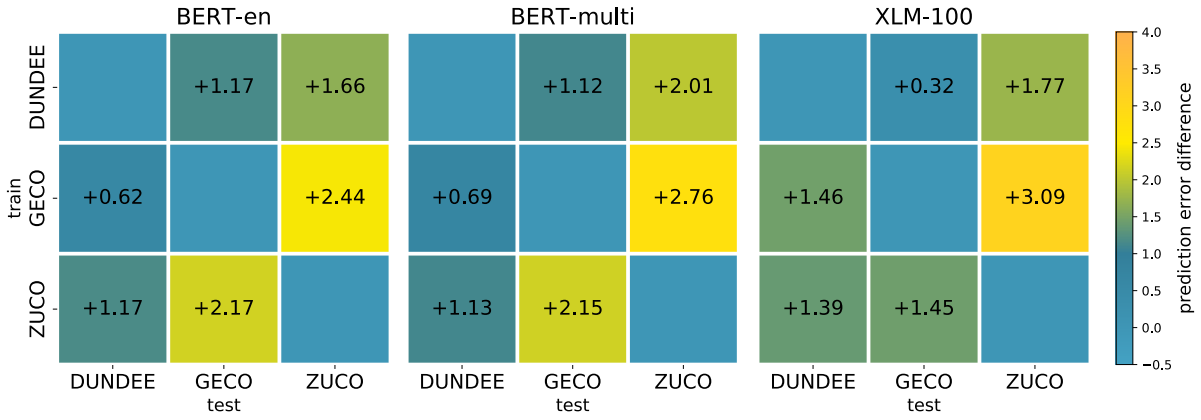


Figure 5.5: Cross-domain evaluation on pre-trained English models. The results are expressed as the difference in the prediction error compared to the in-domain prediction. A smaller error (i.e., a color more similar to the color of the diagonal) represents better domain adaptation.

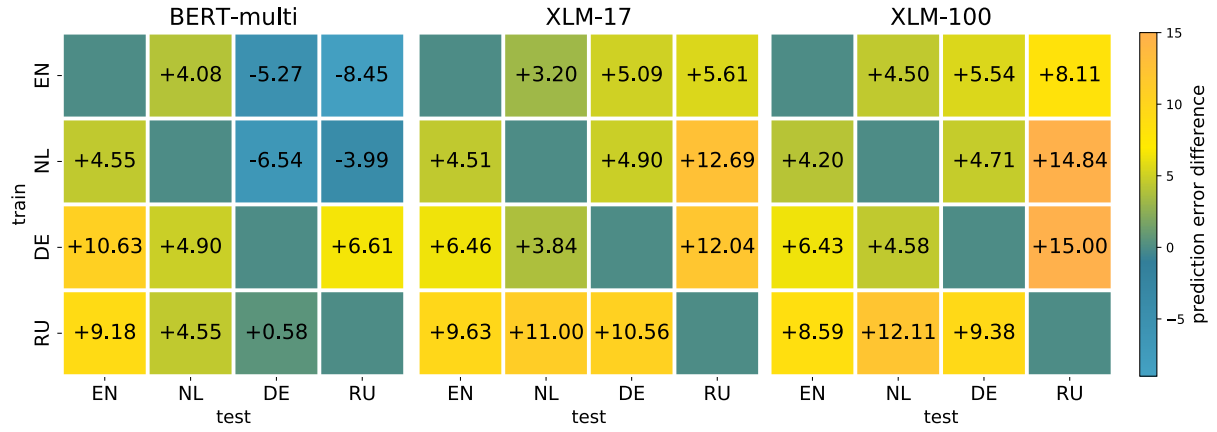


Figure 5.6: Cross-language evaluation on multilingual models across English, Dutch, German and Russian data. The results are expressed as the difference in the prediction error compared to the prediction on the same language. A smaller error (i.e., a color more similar to the color of the diagonal) represents better language transfer.

Cross-Language Evaluation

Figure 5.6 shows the results for cross-language evaluation to probe the language transfer capabilities of the multilingual models. We test models fine-tuned on language A on the test set of language B. It can be seen that BERT-multi generalizes better across languages than the XLM models. This might be due to the fact that the multilingual BERT model is trained on one large

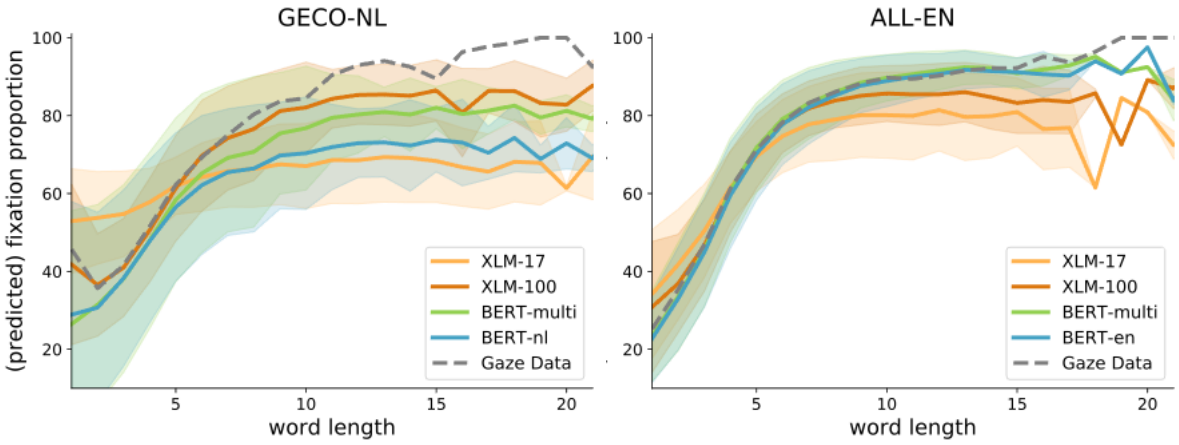


Figure 5.7: Predicted fixation proportion (FPROP) with respect to word length for Dutch and English. The shaded areas show the standard deviation.

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

Eye Tracking Input Properties

Gaze patterns are strongly correlated with word length. Figure 5.7 shows that the fine-tuned models accurately learn to predict higher fixation proportions for longer words. We observe that the predictions of the XLM-100 model follow the trend in the original data most accurately. Similar patterns emerge for the other languages (see Appendix A.3). Notably, the pre-trained models before fine-tuning do not reflect the word length effect.

On the sentence level, we hypothesize that eye tracking features are easier to predict for sentences with higher readability. Figure 5.8 shows the accuracy of the language models for predicting the number of fixations (NFIX) in a sentence relative to the Flesch reading ease score. Interestingly, the pre-trained models without fine-tuning conform to the expected behavior and

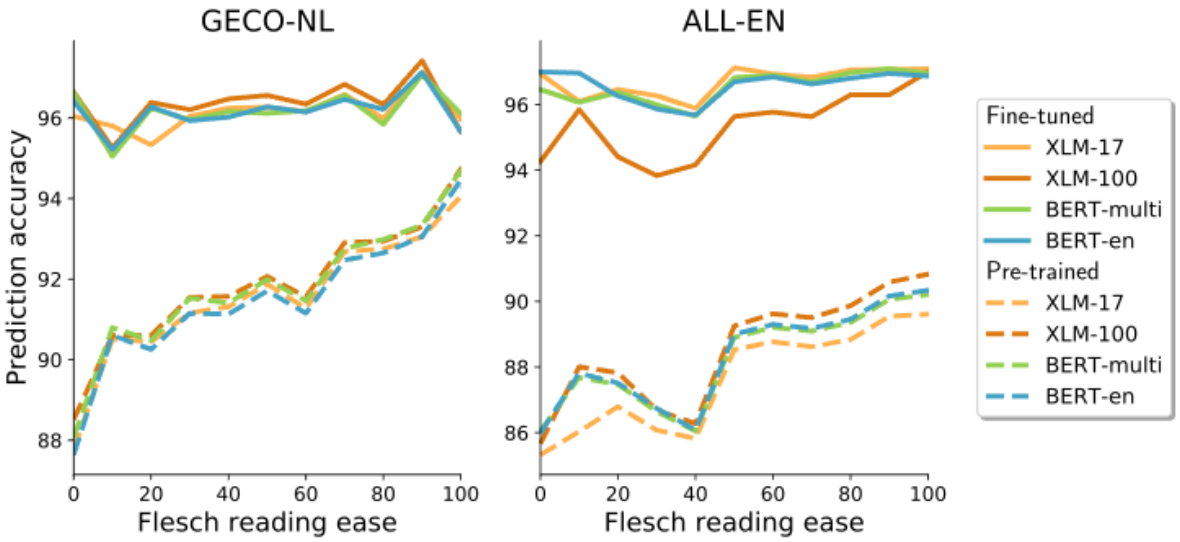


Figure 5.8: Prediction accuracy for the number of fixations (NFIX) in a sentence relative to the Flesch reading ease score of the sentence. A higher Flesch score indicates that a sentence is easier to read. The dashed lines show the results of the pre-trained language models without fine-tuning on eye tracking data.

show a consistent increase in accuracy for sentences with a higher reading ease score. After fine-tuning on eye-tracking data, this behavior is not as visible anymore since the language models achieve constantly high accuracy independent of the readability of the sentences. These results might be explained by the nature of the Flesch readability score, which is based only on the structural complexity of the text (see Appendix A.3 for a description of the Flesch Reading Ease score). Our results indicate that language models trained purely on textual input are more calibrated towards such structural characteristics, i.e., the number of syllables in a word and the number of words in a sentences. In future work, comparing eye movement patterns and text difficulty should rely on readability measures that take into account lexical, semantic, syntactic, and discourse features. This might reveal different patterns between pre-trained and fine-tuned models.

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

5.4 Summary

While the benefits of pre-trained transformer language models have been established, we have yet to understand to which extent these models incorporate human language processing behavior. We take a step in this direction by fine-tuning language models on eye tracking data to predict human reading behavior. Previous work on psycholinguistic modelling finds good fits between cognitive signals and transformer language models. However, the direct prediction of these features has not been explored previously. Our analyses provide insights into the differences between human processing strategies and computational models.

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

The ability of transformer models to achieve such high results in modelling human sentence processing indicates that we can learn more about the cognitive plausibility of these models by predicting behavioral metrics. With this study, we also want to encourage more multilingual research in this field to counteract the anglocentricity in both reading studies and natural language processing (Bender, 2018; Share, 2008).

Finally, the proposed approach for eye tracking prediction also provides a method to eliminate the data scarcity problem when integrating human gaze data in neural networks for NLP (Schwartz et al., 2019; Sood et al., 2020b). As described in Chapter 3, many natural language processing models have shown improvements thanks to eye tracking data on English text. It remains to be seen if the language models fine-tuned on eye movement data achieve similar improvements on English and on other languages.

In the following chapter, we use this approach of finding alignments between computational language models and human cognitive processing signals to evaluate a large range of word embeddings not only leveraging eye tracking data, but also including brain activity data from electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) recordings.

Chapter 6

Evaluating Word Embeddings with Cognitive Processing Signals

Current state-of-the-art machine learning algorithms for language understanding are still mostly black boxes. Word representations are low-dimensional vectors representing the meaning of a word or sentence. The link between a vector of numbers and a humanly interpretable representation of semantics is hidden. This means we cannot comprehend or track the decision-making process of NLP models. Interpretability is key for many NLP applications in order to be able to understand the algorithms' decisions. Moreover, one of the challenges for computational linguistics is to build cognitively plausible models of language processing, i.e., models that integrate multiple aspects of human language processing at the syntactic and semantic level (Keller, 2010). Evaluating and comparing the quality of different word representations is a well-known, largely open challenge. In Chapter 5, we investigated to what extent contextualized language models entail human eye movement patterns during reading. However, there is a need for a more comprehensive multi-modal and multilingual evaluation framework based on comparing NLP models to human language processing behavior. This will contribute to achieve higher cognitive plausibility and better human-grounded interpretability in our computational models.

The contents of this chapter are largely based on the following publications: (1) Hollenstein et al. (2019b). CogniVal: A Framework for Cognitive Word Embedding Evaluation. *CoNLL*. (2) Hollenstein et al. (2020b). CogniVal in Action: An Interface for Customizable Cognitive Word Embedding Evaluation. *COLING*.

This can be achieved through cognitive lexical semantics, a theory that proposes that words are defined by how they are organized in the brain (Miller and Fellbaum, 1992). Huth et al. (2016) showed in a neuroscience study that words are represented in semantic maps across the brain. Recordings of brain activity and gaze patterns play a central role in furthering our understanding of human language and can be used to inspect certain linguistic characteristics in NLP models. Moreover, language representations tuned on brain activity show improved performance on NLP tasks (Schwartz et al., 2019). Hence, it seems natural to evaluate language models against human language processing data, as we propose in this chapter.

In this chapter, we develop *CogniVal*, a framework for cognitive word embedding evaluation. We first present the motivation and design principles of this framework (Section 6.1). We introduce the cognitive data sources and word embeddings (Sections 6.2 and 6.3), as well as the methodology and results of our approach (Sections 6.4 and 6.5). In Section 6.6, we describe the command line interface for CogniVal.

6.1 Background

A promising method for evaluating word representations is by how much they reflect the semantic representations in the human brain. However, most, if not all, previous studies only focus on small datasets and a single modality. In this work, we present the first multi-modal framework for evaluating English word representations based on cognitive lexical semantics. Six types of word embeddings are evaluated by fitting them to 15 datasets of eye tracking, EEG, and fMRI signals recorded during language processing. To achieve a global score over all evaluation hypotheses, we apply statistical significance testing accounting for the multiple comparisons problem. This framework is easily extensible and available to include other intrinsic and extrinsic evaluation methods. We find strong correlations in the results between cognitive datasets, across recording modalities and with their performance on extrinsic NLP tasks.

Word embeddings are the corner stones of state-of-the-art NLP models. Distributional representations, which interpret words, phrases, and sentences as high-dimensional vectors in semantic space, have become increasingly popular. These vectors are obtained by training language models on large corpora to encode contextual information. Each vector represents the meaning of a word. Evaluating and comparing the quality of different word embeddings is an open challenge. Currently, word embeddings are evaluated with extrinsic or intrinsic methods. Extrinsic

evaluation is the process of assessing the quality of the embeddings based on their performance on downstream NLP tasks, such as question answering or entity recognition. However, embeddings can be trained and fine-tuned for specific tasks, but this does not mean that they accurately reflect the meaning of words.

Instead, intrinsic evaluation methods, such as word similarity and word analogy tasks, merely test single linguistic aspects. These tasks are based on conscious human judgements. Conscious judgements can be biased by subjective factors and the tasks themselves might also be biased (Nissim et al., 2020). Additionally, the correlation between intrinsic and extrinsic metrics is not very clear, as intrinsic evaluation results fail to predict extrinsic performance (Chiu et al., 2016; Gladkova and Drozd, 2016). Finally, both intrinsic and extrinsic evaluation types often lack statistical significance testing and do not provide a global quality score.

In this work, we focus on the *intrinsic subconscious evaluation method* (Bakarov, 2018b), *which evaluates English word embeddings against the lexical representations of words in the human brain, recorded when passively understanding language*. Cognitive lexical semantics proposes that words are defined by how they are organized in the brain (Miller and Fellbaum, 1992). As a result, brain activity data recorded from humans processing language is arguably the most accurate mental lexical representation available (Søgaard, 2016). Recordings of brain activity play a central role in furthering our understanding of how human language works. To accurately encode the semantics of words, we believe that embeddings should reflect this mental lexical representation.

Evaluating word embeddings with cognitive language processing data has been proposed previously. However, the available datasets are not large enough for powerful machine learning models, the recording technologies produce noisy data, and most importantly, only a few datasets are publicly available. Furthermore, since brain activity and eye tracking data are generally very noisy, correlating distances between representations does not provide sufficient statistical power to compare embedding types (Frank, 2017). For this reason we evaluate the embeddings by exploring how well they can predict human processing data. We build on Søgaard (2016)'s theory of evaluating embeddings with this task-independent approach based on cognitive lexical semantics and examine its effectiveness. The design of our framework is based on three principles:

1. **Multi-modality**: Evaluate against various modalities of recording human signals to counteract the noisiness of the data.

2. **Diversity** within modalities: Evaluate against different datasets within one modality to make sure the number of samples is as large as possible.
3. **Correlation** of results should be evident across modalities and even between datasets of the same modality.

The contributions in this chapter are as follows. We describe and test CogniVal, the first framework of cognitive word embedding evaluation to follow these principles. We evaluate different embedding types against a combination of 15 cognitive data sources, acquired with three modalities: eye tracking, electroencephalography (EEG), and functional magnetic resonance imaging (fMRI). The word representations are evaluated by assessing their ability of predicting cognitive language processing data (see Figure 6.1). Our main findings when evaluating six state-of-the-art word embeddings with CogniVal show that the majority of embedding types significantly outperform a baseline of random embeddings when predicting a wide range of cognitive features. Additionally, the results show consistent correlations between datasets of the same modality and across different modalities, validating the hypothesis of our approach. Finally, we present an exploratory but promising correlation analysis between the scores obtained using our intrinsic evaluation methods and the performance on extrinsic NLP tasks.

Related Work

Mitchell et al. (2008) pioneered the use of word embeddings to predict patterns of neural activation when subjects are exposed to isolated word stimuli. More recently, this dataset and other fMRI resources have been used to evaluate learned word representations. For instance, Abnar et al. (2018) and Rodrigues et al. (2018) evaluate different embeddings by predicting the neuronal activity from the 60 nouns presented by Mitchell et al. (2008). Søgaard (2016) shows preliminary results in evaluating embeddings against continuous text stimuli in eye tracking and fMRI data. Moreover, Beinborn et al. (2019) recently presented an extensive set of language–brain encoding experiments. Specifically, they evaluated the ability of an ELMo language model to predict brain responses of multiple fMRI datasets. EEG data has been used for similar purposes. Schwartz and Mitchell (2019) and Ettinger et al. (2016) show that components of event-related potentials can successfully be predicted with neural network models and word embeddings.

However, these approaches mostly focus on one modality of brain activity data from small individual cognitive datasets. The lack of data sources has been one reason why this type of

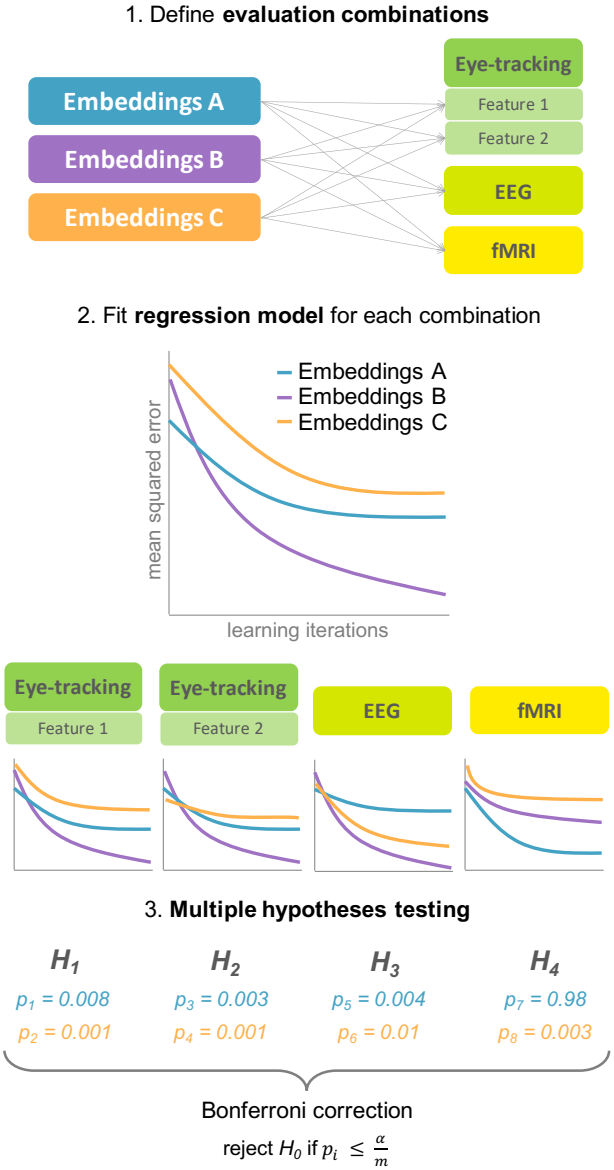


Figure 6.1: Overview of the cognitive word embedding evaluation process of CogniVal.

evaluation has not been too popular until now (Bakarov, 2018a). Hence, in this work we collected a wide range of cognitive data sources including eye tracking, EEG and fMRI. This will ensure coverage of different features, and consequently of the cognitive processing taking place in the human brain during reading.

Evidence from cognitive neuroscience Murphy et al. (2018) review computational approaches to the study of language with neuroimaging data and show how different types of words activate

neurons in different brain regions. Similarly, mapping fMRI data from subjects listening to stories to the activated brain regions revealed semantic maps of how words are distributed across the human cerebral cortex (Huth et al., 2016). Furthermore, word predictability and semantic similarity show distinct patterns of brain activity during language comprehension: semantic distances can have neurally distinguishable effects during language comprehension (Frank and Willems, 2017). These findings support the theory that brain activity data reflects lexical semantics and is thus an appropriate foundation for evaluating the quality of word embeddings.

6.2 Word Representations

Pre-trained language models are essential components in state-of-the-art NLP systems. Word embeddings or word representations are simply vectors of hidden states of these trained language models. We chose six commonly used pre-trained embeddings to evaluate against the cognitive data sources. See Table 6.1 for an overview of the dimensions of each embedding type. We evaluate the following types of word embeddings:

- **Glove**: Embeddings of different dimensions trained on aggregated global word-word co-occurrence statistics over a corpus of 6 billion words (Pennington et al., 2014).
- **Word2vec**: Non-contextual embeddings trained on 100 billion words from a Google News dataset (Mikolov et al., 2013b).
- **WordNet2Vec**: Embeddings representing the conversion from semantic networks into semantic spaces (Saedi et al., 2018). Trained on WordNet, a lexical ontology for English that comprises over 155,000 lemmas (but trained only on 60,000 words).
- **FastText**: Pre-trained embeddings using character n-grams to compose the vector of the full words (Mikolov et al., 2018). We evaluate the embeddings with and without subword information trained on 16 billion tokens of Wikipedia sentences as well as the ones trained on 600 billion tokens of Common Crawl.
- **ELMo**: Embeddings modeling both complex characteristics of word use (i.e., syntax and semantics), and how these uses vary across linguistic contexts (Peters et al., 2018). These word vectors are learned functions of the internal states of a deep bidirectional language

Embeddings	Dimension	Hidden Layer Units
Glove	50	[30, 26, 20, 5]
Glove	100	[50, 30]
Glove	200	[100, 50]
Glove	300	[150, 50]
Word2vec	300	[150, 50]
WordNet2vec	850	[400, 200]
FastText	300	[150, 50]
ELMo	1024	[600, 200]
BERT	768	[400, 200]
BERT	1024	[600, 200]

Table 6.1: Overview of word embeddings evaluated with CogniVal. The last column shows the search space of the grid search for the number of units in the hidden layer.

model, which is pre-trained on a large text corpus. We take the first of the three output layers, containing the context insensitive word representations.

- **BERT**: Contextual bidirectional word representations, based on the idea that fine-tuning a pre-trained language model can help the model achieve better results in the downstream tasks (Devlin et al., 2019). We take the hidden states of the second to last of the 12 output layers as the representation for each token.

6.3 Cognitive Data Sources

In this work, we consider three modalities of recording cognitive language processing signals: eye tracking, electroencephalography (EEG), and functional magnetic resonance imaging (fMRI). All three methods are complementary in terms of temporal and spatial resolution as well

[1] https://www.kilgarriff.co.uk/bnc-readme.html

	Data source	Stimulus	Subj.	Tokens	Types	Coverage
Eye Tracking	GECO (Cop et al., 2017)	text	14	68606	5383	95%
	Dundee (Kennedy et al., 2003)	text	10	58598	9131	94%
	CFILT-Sarcasm (Mishra et al., 2016a)	text	5	23466	4237	85%
	ZuCo (Hollenstein et al., 2018)	text	12	13717	4384	90%
	CFILT-Scanpath (Mishra et al., 2017b)	text	5	3677	1314	89%
	Provo (Luke and Christianson, 2017)	text	84	2743	1192	95%
	UCL (Frank et al., 2013)	text	43	1886	711	98%
	all eye tracking (aggregated)	text	-	26353	16419	88%
EEG	ZuCo (Hollenstein et al., 2018)	text	12	13717	4384	90%
	Natural Speech (Broderick et al., 2018)	speech	19	12000	1625	98%
	UCL (Frank et al., 2015)	text	24	1931	711	98%
	N400 (Broderick et al., 2018)	text	9	150	140	100%
fMRI	Harry Potter (Wehbe et al., 2014)	text	8	5169	1295	92%
	Alice (Brennan et al., 2016)	speech	27	2066	588	99%
	Pereira (Pereira et al., 2018)	text/image	15	180	180	99%
	Nouns (Mitchell et al., 2008)	image	9	60	60	100%

Table 6.2: Cognitive data sources used in this work. Coverage is the percentage of the vocabulary in each data source that occurs in the British National Corpus list of the most frequent English words.[1]

as the directness in the measurement of neural activity (Mulert, 2013). For the word embedding evaluation, we selected a wide range of datasets from these three modalities to ensure a more diverse and accurate representation of the brain activity during language processing.

Table 6.2 shows an overview of the cognitive data sources used, which are described in more detail below. Since the processing in the brain differs depending on whether the information is accessed via the visual or auditory system (Price, 2012), we include data of different stimuli, e.g., participants reading sentences or listening to audiobooks. Moreover, our collection of cognitive data sources contains datasets of both isolated (single words) and continuous (words in context, i.e., sentences or stories) stimuli. All datasets include English language stimuli and the participants were native speakers or highly proficient.

Eye Tracking

Eye tracking is an indirect measure of cognitive activity. Gaze patterns are highly correlated with the cognitive load associated with different stages of human text processing (Rayner, 1998). For instance, fixation duration is higher for long, infrequent and unfamiliar words (Just and Carpenter, 1980). Section 3.2.1 covers the psycholinguistic research on eye movements during reading in more depth.

All eye tracking datasets used in this work were recorded from natural, self-paced reading. By using as many features as available from each dataset, including features characterizing basic, early and late word processing aspects, the goal is to cover the whole language understanding process at word level. Each dataset provides different eye tracking features. The most common features, available in all 7 datasets are: first fixation duration, first pass duration, mean fixation duration, total fixation duration and number of fixations. For a complete list and description of the eye tracking features available in each corpus see Table 6.3.

Eye tracking preprocessing Gaze vectors consist of specific features, which are extracted from the reading times, fixations and regressions on each word. Feature values are aggregated on word type level and scaled between 0 and 1. The feature values were averaged over all subjects within a dataset. This preprocessing step is done separately for each data source before combining them. We show that combining gaze data from different sources can be helpful for NLP applications, even when they are recorded with different devices and filtering methods. We scale each data source individually and then aggregate over word types, so that we can analyze the prediction performance of word embeddings for single or aggregated eye tracking features.

Electroencephalography (EEG)

Electroencephalography records electrical activity from the brain by measuring the voltage fluctuations through the scalp with high temporal resolution. Hauk and Pulvermüller (2004) present evidence for the modulation of early electrophysiological brain responses by word frequency. This supports the theory that lexical access from written word stimuli is an early process that follows stimulus presentation by less than 200 ms.

The EEG datasets used in this work were either recorded from reading sentences or listening to natural speech. Word-level brain activity could be extracted by either mapping it to eye tracking

Description	Data source
First fixation duration	Dundee, GECO, Provo, UCL, ZuCo
First pass duration (first fixation duration in the first pass reading)	Dundee, GECO, Provo, UCL, ZuCo
Mean fixation duration	Dundee, GECO, Provo, ZuCo, CFILT-Sarcasm, CFILT-Scanpath
Fixation probability	Dundee
Re-read probability	Dundee
Total fixation duration	Dundee, GECO, Provo, ZuCo
Total duration of all regression going from this word	Dundee
Total duration of all regression going to this word	Dundee
Number of fixations	Dundee, GECO, Provo, ZuCo
Number of long regression (>3 tokens) going from this word	Dundee
Number of long regression (>3 tokens) going to this word	Dundee
Number of re-fixations	Dundee
Number of regressions going from this word	Dundee, Provo
Number of regressions going to this word	Dundee, Provo
Duration of the last fixation on the current word	GECO
Go-past time	GECO, Provo, UCL, ZuCo
No fixation occurred in first-pass reading	GECO, Provo
Right-bounded reading time	UCL

Table 6.3: Eye tracking features provided in the gaze corpora used for CogniVal.

cues (ZuCo) or auditory triggers (Natural Speech), or by recording only the last word in each sentence (N400), or through serial presentation of the words (UCL). Standard preprocessing steps for EEG data, including band-pass filtering and artifact removal, are performed in the same manner for all four data sources.

EEG preprocessing All four EEG datasets are converted to the EEGLab format,[2] if not already provided in this format. The UCL dataset was preprocessed by the authors. For the other three datasets, bandpass filtering, artifact removal (i.e., removing blinks and other muscle activity), and quality assessment was performed with Automagic.[3] Only channels marked as 'good quality' by Automagic were retained.

After preprocessing and retaining only the subjects with good data quality, we use the data of 3 subjects from the N400 dataset, 14 subjects from Natural Speech, 12 subject from ZuCo and the same number of subjects as originally from UCL (i.e., 24). The EEG data is aggregated over all available subjects and over all occurrences of a token. This yields an n-dimensional vector, where n is the number of electrodes, ranging from 32 to 130, depending on the EEG device used to record the data. EEG data can be aggregated over all subjects within one dataset, because the number and locations of electrodes are identical. However, due to the differences in the number of electrodes between datasets, we cannot aggregate over all EEG datasets.

Functional Magnetic Resonance Imaging (fMRI)

Functional magnetic resonance imaging is a technique for measuring and mapping brain activity by detecting changes associated with blood flow. fMRI has a temporal resolution of two seconds, which means that with continuous stimuli such as natural reading or story listening, one scan covers multiple words. We use datasets of isolated stimuli (e.g., the Nouns dataset) as well as continuous stimuli (e.g., Harry Potter). While it is easier to extract word-level signals from isolated stimuli, continuous stimuli allow extracting signals in context over a wider vocabulary.

Where multiple trials were available, the brain activation for each word is calculated by taking the mean over the scans. Moreover, if the stimulus is continuous (Harry Potter and Alice datasets), the data is aligned with an offset of four seconds to account for the hemodynamic delay.[4]

[2] `https://sccn.ucsd.edu/eeglab/index.php`

[3] `https://github.com/methlabUZH/automagic`

[4] The fMRI signal measures a brain response to a stimulus with a delay of a few seconds, and it decays slowly over a duration of several seconds (Miezin et al., 2000). For continuous stimuli, this means that the response to previous stimuli will have an influence on the current signal. Thus, context of the previous words is taken into account.

FMRI preprocessing FMRI data contains representations of neural activity of millimeter-sized cubes called voxels. Standard fMRI preprocessing methods such as motion correction, slice timing correction, and co-registration had already been applied before. To select the voxels to be predicted we use the pipeline provided by Beinborn et al. (2019) to read the fMRI data, align the scans, and select the voxels. This pipeline consists of extracting corresponding scan(s) for each word, and randomly selecting 100, 500 and 1000 voxels for the Harry Potter, Pereira and Nouns datasets, respectively. The published version of the Alice dataset provided the preprocessed signal averaged for six regions of interest, hence for this particular dataset we predict the activation only for these regions. We used the Nouns and Pereira readers as is and modified the Harry Potter and Alice readers to extract word-level signals. Finally, the fMRI data is converted to n-dimensional vectors, where n is the number of randomly selected voxels (100, 500 or 1000) or regions (6).

6.4 Embedding Evaluation Method

In order to evaluate the word embeddings against human lexical representations, we fit the embeddings to a wide range of cognitive features, i.e., eye tracking features and activation levels of EEG and fMRI. This section describes how these models were trained and evaluated. After evaluating each combination separately, we test for statistical significance taking into account the multiple comparisons problem. See Figure 6.1 for an overview of the evaluation process.

6.4.1 Regression Models

We fit neural regression models to map word embeddings to cognitive data sources. Predicting multiple features from different sources and modalities allows us to evaluate different aspects of capturing the semantics of a word. Hence, separate models are trained for all combinations. For instance, fitting FastText embeddings to EEG vectors from ZuCo, or fitting ELMo embeddings to the first fixation durations of the Dundee corpus.

For the regression models, we train neural networks with k input dimensions, one dense hidden layer of n nodes using ReLU activations and an output layer of m nodes using linear activation. The model is a multiple regression with layers of dimensions k-n-m, where k is the number of

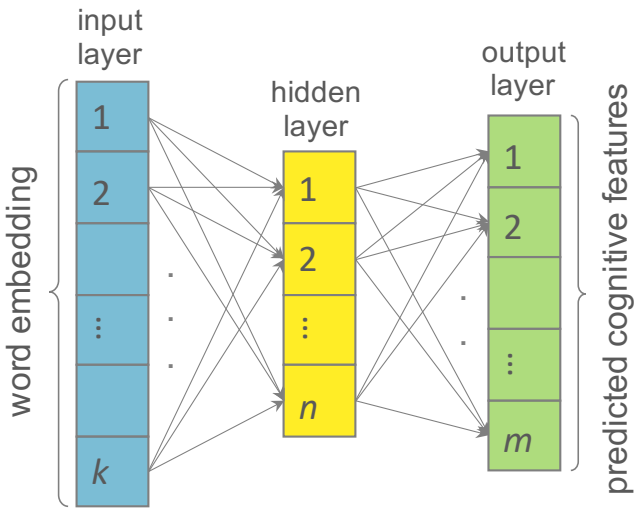


Figure 6.2: Neural architecture of the regression models. The neural networks are trained with k input dimensions, one dense hidden layer of n nodes using ReLU activation and an output layer of m nodes using linear activation. The model is a multiple regression with layers of dimension k-n-m.

dimensions of the word embeddings and m changes depending on the cognitive data source to be predicted (Figure 6.2). For predicting single eye tracking features m equals 1, whereas for EEG or fMRI vectors m is the dimension of the cognitive data vector, representing the number of electrodes in the EEG data or the number of voxels in the fMRI data. The loss function optimizes the mean squared error (MSE) and uses an Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.001.

5-fold cross-validation is performed for each model (80% training data and 20% test data). The optimal number of nodes n in the hidden layer is selected individually for each combination of cognitive data source and embedding type. To this end, a grid search is performed before training, which is evaluated on a validation set consisting of 20% of the training data with 3-fold cross validation (see Table 6.1 for details on the search space). The best model is then saved and used to predict the cognitive feature for each word in the test set. Finally, the results are measured with the mean squared error, averaged over all predicted words.

CogniVal allows for evaluation against another word embedding type as well as evaluation against a random baseline. To generate a fair baseline we create random vectors for each word of n dimensions, corresponding to the same number of dimensions of the embeddings to be evaluated.

6.4.2 Multiple Hypotheses Testing

With the purpose of achieving consistency and going towards a global quality metric that can be combined with other evaluation methods, we perform statistical significance testing on each hypothesis. There is one hypothesis for each evaluation combination, i.e., of an embedding type and a cognitive data source. The hypothesis is that the word embeddings can predict the features from the cognitive data source more accurately than the random baseline.

Since the distribution of our test data is unknown and the datasets are small, we perform a Wilcoxon signed-rank test for each hypothesis (Dror et al., 2018). Additionally, we face a multiple hypothesis testing problem, which occurs when a number of individual hypothesis tests are considered simultaneously. In this case, the significance or the error rate of individual tests no longer represents the error rate of the combined set of tests. Multiple hypothesis testing methods correct error rates for this issue (Rupert Jr et al., 2012). To counteract this multiple hypotheses problem by controlling the occurrence of false positives, we apply the conservative Bonferroni correction (Bonferroni, 1936), which controls for the family-wise error rate (FWER). Under the Bonferroni correction, the global null hypothesis is rejected if $p < \alpha/N$, where N is the number of hypotheses (Dror et al., 2017). In our setting, $\alpha = 0.01$ and $N = 4$ for EEG (one hypothesis per EEG data source), $N = 59$ for for fMRI (one hypothesis per participant of each fMRI data source), and $N = 42$ for eye tracking (one hypothesis per feature per eye tracking corpus). The significance ratios are shown in Figure 6.3.

This approach of significance testing can easily be used in combination with other intrinsic and extrinsic evaluation methods. However, with additional external results, the number of hypotheses will increase. Then it would be more sensible to apply a Benjamini-Hochberg correction, which introduces the idea of a false discovery rate (FDR) to control for multiple hypotheses testing (Benjamini and Hochberg, 1995). Controlling the FDR instead of the FWER is less stringent and would allow for a more sensible correction as the number of hypotheses increases.

6.5 Results & Discussion

In this section we present the results for our CogniVal framework. First, we discuss the prediction results for the different word embedding types. We also present additional analyses on the

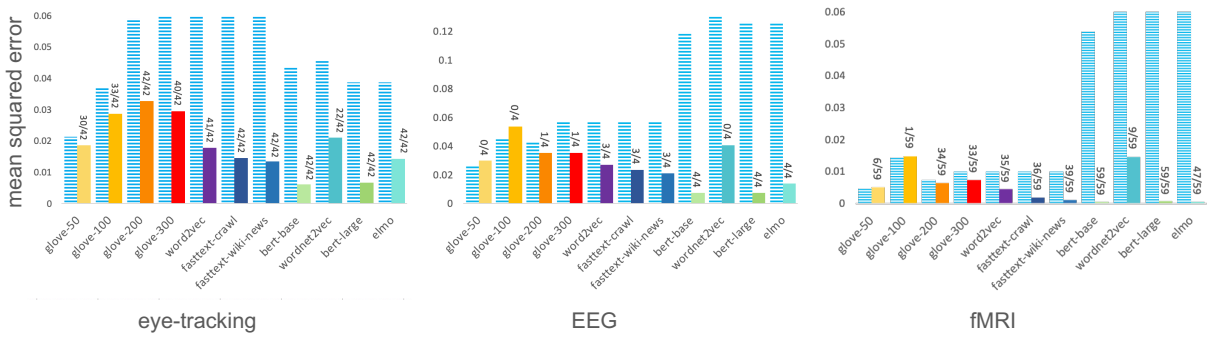


Figure 6.3: Aggregated results for all embeddings predicting cognitive features for all datasets of a modality. The results are sorted by dimension of embeddings in increasing order from left to right. The striped blued bars represent random baseline. The labels on the embedding bars show the ratio of significant results under the Bonferroni correction to the total number of hypotheses.

design decisions made for the preprocessing and feature extraction of the cognitive data sources. Finally, we analyze the correlation of the results, across and within cognitive data modalities as well as a comparison of the CogniVal results and other reported extrinsic results.

Prediction Results

The majority of the word embedding types are able to predict eye tracking features, EEG and fMRI vectors significantly better than random baseline (Figure 6.3). BERT, ELMo and FastText embeddings achieve the best predictions. The detailed results can be found in Tables 6.4, 6.5 and 6.6 for eye tracking, fMRI, and EEG datasets, respectively. The random baseline can be considered a rather naive choice, but this setting also allows us to compare the performance between word embedding types.

When predicting single eye tracking features, the performance varies greatly. For instance, the prediction error on the number of fixations and total reading time from the ZuCo dataset is much lower than for the first fixation duration (Table 6.7). This suggests that more general eye tracking features covering the complete reading process of a word are easier to predict.

In the case of predicting voxel vectors of fMRI data, the results improve when choosing a larger number of voxels (see Table 6.8). Hence, in the remainder of this work we present only the results for 1000 voxels.

We also examined the EEG results in more depth. We analyze which electrodes are predicted more accurately and which electrode values are very difficult to predict by the best-performing

Embeddings	GECO	ZuCo	Provo	Dundee	Sarcasm	Scanpath	UCL
Glove-50	0.010	0.008	0.031	0.010	0.016	0.023	0.044
Glove-100	0.018	0.017	0.051	0.014	0.027	0.043	0.054
Glove-200	0.026	0.024	0.047	0.021	0.039	0.038	0.054
Glove-300	0.020	0.019	0.047	0.016	0.033	0.038	0.059
Word2vec	0.015	0.011	0.024	0.014	0.022	0.018	0.028
Fasttext-Crawl	0.011	0.009	0.017	0.014	0.020	0.010	0.023
Fasttext-WikiNews	0.010	0.008	0.015	0.014	0.019	0.009	0.019
Wordnet2vec	0.018	0.012	0.027	0.016	0.023	0.017	0.040
Bert-Base	0.007	0.003	0.006	0.008	0.009	0.006	0.003
Bert-Large	0.008	0.004	0.006	0.008	0.011	0.006	0.003
Elmo	0.012	0.009	0.020	0.011	0.014	0.012	0.021

Table 6.4: Absolute mean squared error averaged over all features for each combination, i.e., the averaged error of all eye tracking features for each dataset.

embeddings (BERT) and aggregated over all embedding types. This is exemplified by Figure 6.4, which shows the 20 best and worst predicted electrodes of the ZuCo dataset for the BERT embeddings of 1024 dimensions as well as aggregated over all embedding types. The middle central electrodes are predicted more accurately. The middle central electrodes are known to register the activity of the Perisylvian cortex, which is relevant for language-related processing (Catani et al., 2005). Moreover it can be speculated that there is a frontal asymmetry between the electrodes on the left and right hemispheres.

Cognitive Data Implications

The diversity of cognitive data sources chosen for this work allows us to analyze and compare results on several levels and between several cognitive metrics. In order to conduct this evaluation on a collection of 15 datasets from three modalities, many crucial decisions were taken about preprocessing, feature extraction and evaluation type. Since there are different methods for processing different types of cognitive language understanding signal, it is important to make these decisions transparent and reproducible. It is a challenge to segment brain activity data correctly and meaningfully into word-level signals from naturalistic, continuous language stimuli (Hamilton and Huth, 2018). This makes consistent preprocessing across data sources even more

Embeddings	Harry Potter	Nouns	Alice	Pereira
Glove-50	0.005	0.204	0.036	0.044
Glove-100	0.015	0.220	0.069	0.055
Glove-200	0.007	0.224	0.036	0.050
Glove-300	0.008	0.224	0.038	0.050
Word2vec	0.005	0.209	0.010	0.044
Fasttext-Crawl	0.002	0.194	0.009	0.039
Fasttext-Wikinews	0.001	0.185	0.004	0.037
Wordnet2vec	0.015	0.203	0.050	0.046
Bert-Base	0.001	0.042	0.001	0.012
Bert-Large	0.001	0.055	0.001	0.013
Elmo	0.001	0.135	0.001	0.034

Table 6.5: Absolute mean squared error averaged over all voxels in each fMRI dataset.

Embeddings	N400	Natural Speech	ZuCo	UCL
Glove-50	0.067	0.014	0.009	0.030
Glove-100	0.126	0.023	0.011	0.056
Glove-200	0.071	0.017	0.013	0.040
Glove-300	0.067	0.018	0.014	0.043
Word2vec	0.047	0.017	0.012	0.033
Fasttext-Crawl	0.042	0.013	0.011	0.029
Fasttext-WikiNews	0.037	0.012	0.010	0.026
Wordnet2vec	0.089	0.020	0.015	0.005
Bert-Base	0.014	0.005	0.006	0.030
Bert-Large	0.012	0.006	0.006	0.006
Elmo	0.024	0.008	0.008	0.015

Table 6.6: Absolute mean squared error averaged over all electrodes in each EEG dataset.

important.

Another challenge is to consolidate the cognitive features to be predicted. For instance, we chose a wide selection of eye tracking features that cover early and late word processing. However, choosing only general eye tracking features such as total reading time would also be a viable

Embeddings	nFix	TRT	FFD
Glove-300	0.010	0.017	0.027
Word2vec	0.009	0.010	0.016
Fasttext-Crawl	0.008	0.007	0.012
Bert-Base	0.005	0.003	0.004
Wordnet2vec	0.010	0.010	0.019
Elmo	0.008	0.009	0.011
Mean	**0.008**	**0.009**	**0.015**

Table 6.7: Comparison of the MSE from word embeddings predicting single eye tracking features: number of fixations (nFix), first fixation duration (FFD) and total reading time of a word (TRT).

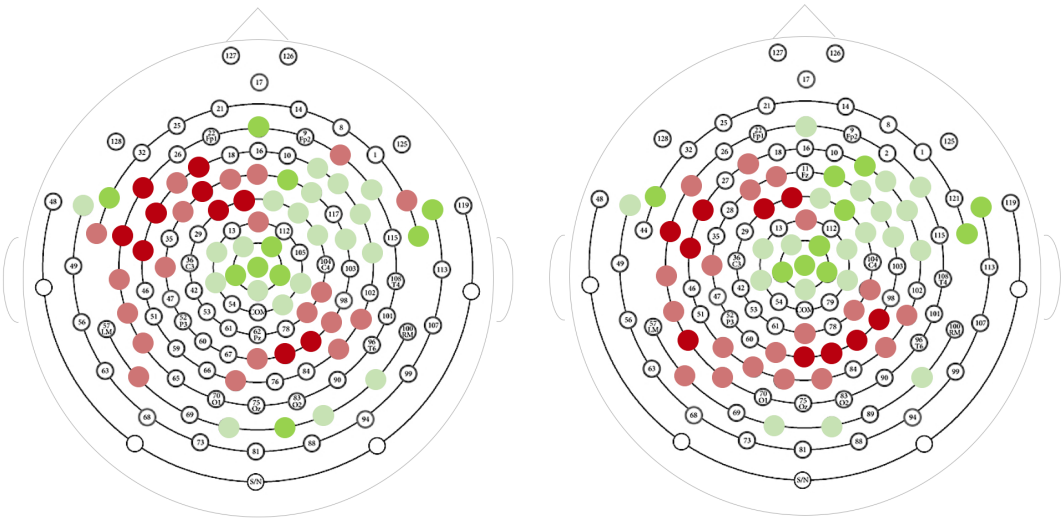
	Voxels		
Embeddings	100	500	1000
Glove-300	0.119	0.081	**0.078**
Word2vec	0.103	**0.075**	**0.075**
Fasttext-Crawl	0.092	0.070	**0.069**
Bert-Base	0.020	0.017	**0.016**
Wordnet2vec	0.105	0.077	**0.076**
Elmo	0.067	0.051	**0.050**

Table 6.8: MSE Results of predicting different amounts of randomly selected voxels from the fMRI datasets.

strategy. On the other hand, the EEG evaluation could be more coarse-grained, one could also try to predict known ERP effects (e.g., Ettinger et al. (2016)) or features selected based on frequency bands. Moreover, the voxel selection in the fMRI preprocessing could be improved by either predicting all voxels or applying information-driven voxel selection methods (Beinborn et al., 2019).

Correlation Between Modalities

Next, we analyze the correlation between the predictions of the three modalities (Figure 6.5). There is a strong correlation between the results of predicting eye tracking, EEG and fMRI



(a) (b)

Figure 6.4: EEG electrode analysis, (a) for BERT (large) and (b) aggregated over all embedding types. Red = worst predicted electrodes, green = best predicted electrodes.

features. This implies that word embeddings are actually predicting brain activity signals and not merely preprocessing artifacts of each modality. Moreover, the same correlation is also evident between individual datasets within the same modality. As an example, Figure 6.6 (left) shows the correlation of the results predicted for the Natural Speech and ZuCo EEG datasets, where the first had speech stimuli and the latter text. Figure 6.6 (right) reveals the same positive correlation for two EEG datasets that were preprocessed differently and were recorded with a different number of electrodes. Moreover, the UCL dataset contains word-by-word reading and the N400 contains natural reading of full sentences. Nevertheless, CogniVal shows a clear correlation between the results on both datasets, meaning that the predictions are independent of specific preprocessing decisions, differences in the experimental design or the amount of noise in different EEG datasets.

Correlation with Extrinsic Evaluation Results

We performed a simple comparison between the results of word embeddings predicting cognitive language processing signals and the performance of the same embedding types in downstream tasks. We collected results for two NLP tasks: on the SQuAD 1.1 dataset for question answering (Rajpurkar et al., 2016) and on the CoNLL-2003 test split for named entity recognition (Tjong Kim Sang and De Meulder, 2003).

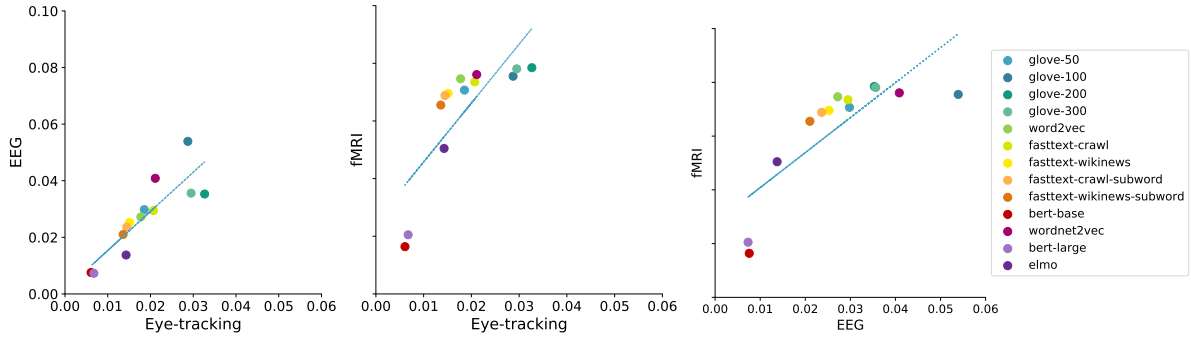


Figure 6.5: Correlation plots showing MSE results between all three modalities of cognitive signals.

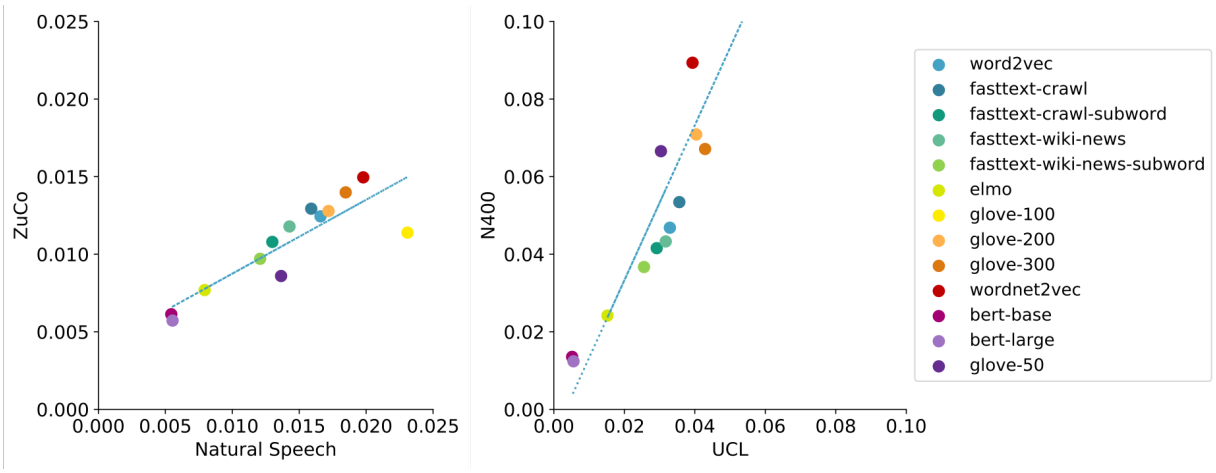


Figure 6.6: Correlation of the MSE results between EEG datasets. On the left, we correlate the mean squared error achieved by the ZuCo dataset to the error achieved by the Natural Speech dataset, on the right between the N400 and the UCL dataset.

The SQuAD results are taken from Devlin et al. (2019) for BERT, from Mikolov et al. (2018) for FastText, and from Peters et al. (2018) for ELMo. The NER results are from the same source for ELMo and BERT, for Glove-50 from Pennington et al. (2014) and for Glove-200 from Ghannay et al. (2016). We correlated these results to the prediction results over all cognitive data sources. Figure 6.7 shows the correlation plots between the CogniVal results and the two downstream tasks.

While this is merely an exploratory analysis, it shows interesting findings: If the cognitive embedding evaluation correlates with the performance of the embeddings in extrinsic evaluation tasks, it might be used not only for evaluation but also as a predictive framework for word embedding model selection. This is especially noteworthy, since it does not seem to be the case for other intrinsic methods (Chiu et al., 2016).

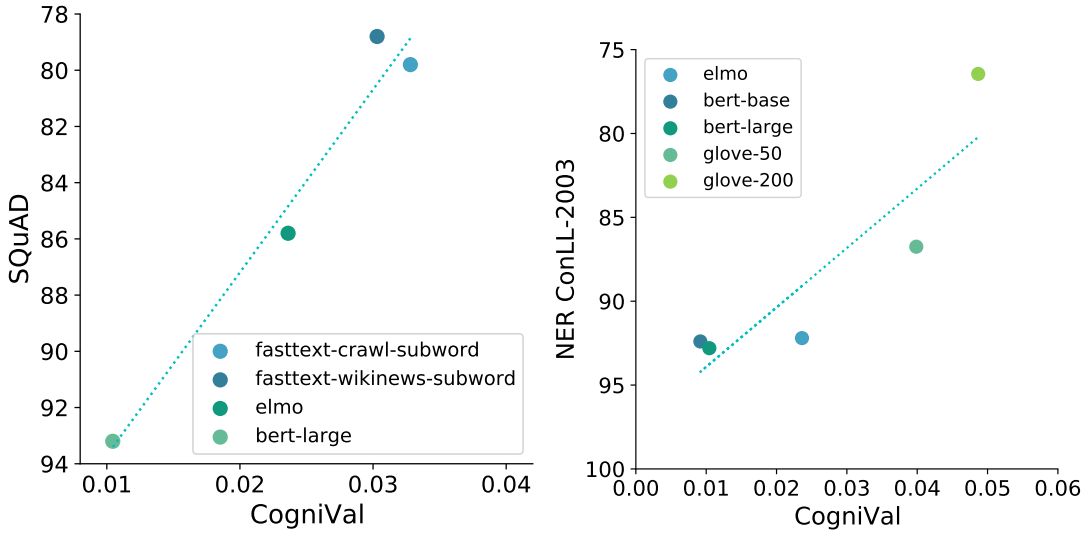


Figure 6.7: Correlation between the SQuAD 1.1 task (left) and NER on CoNLL-2003 (right) and the CogniVal results. The y-axis represent the accuracy of the extrinsic evaluation tasks, and the x-axis the aggregated mean squared error of the CogniVal results.

6.6 CogniVal in Action

The interface presented in this section is based on the CogniVal framework. We presented the first encompassing framework for cognitive word embedding evaluation. We improve and extend the original features of CogniVal and provide a simple command line interface for scalable and customized experiments. We build a user interface for CogniVal to make it usable and accessible for NLP practitioners. The CogniVal command line interface (CLI) offers pre-processed cognitive data sources, readily provided for evaluation in a user-friendly interaction. It supports and complements other intrinsic and extrinsic evaluation methods for word embeddings. The CogniVal CLI is a unified framework, allowing the evaluation of a large set of existing pre-trained embeddings but also of custom embeddings on a large range of cognitive sources.

6.6.1 Command Line Interface

The CogniVal command line interface is openly available and can be easily installed with `pip`.[5] It is implemented in Python (version 3.7.4) and provides an interactive shell using the

[5] `https://github.com/DS3Lab/cognival-cli`

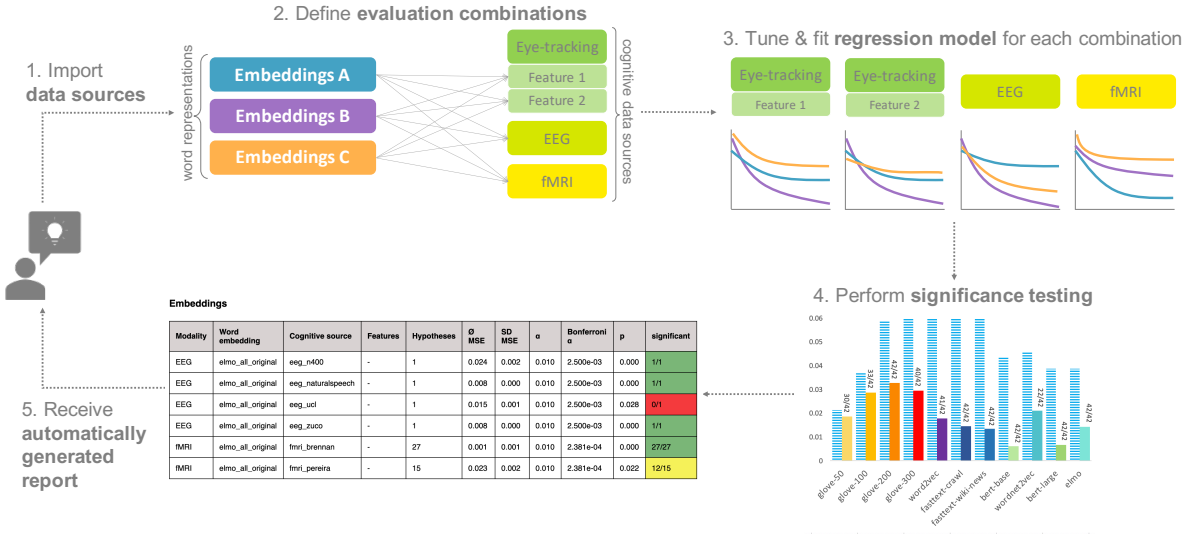


Figure 6.8: Schematic overview of the user interaction process of the CogniVal interface. First, the user can import custom cognitive data sources or word embeddings if needed. Then, the use defines the desired evaluation combinations. CogniVal runs the evaluations and finally returns a report including all results including the significance testing.

`python-nubia` library.[6] As described in Section 6.3, we collected and prepared 15 cognitive data sources and evaluated 6 pre-trained embedding types, including GloVe, word2vec, WordNet2Vec, FastText, ELMo and BERT. The command line interface provides these preprocessed data types.

The evaluation process is automatized in the CogniVal CLI (Figure 6.8). First, the user defines the general evaluation configuration, including path specifications and training parameters (command: `config`). If required, the user can then import custom word representations as well as custom cognitive data sources, using the `import` function. Second, the user specifies the embedding/cognitive-data combinations to be evaluated, as well as the hyper-parameter ranges for the neural regression models. Moreover, if requested, CogniVal generates random vectors of the same dimension of the embeddings to be evaluated. The embeddings can also be evaluated against this random baseline. As an improvement from the method described in Section 6.4, the CogniVal CLI automatically generates 10 sets of different random embeddings and averages over the results for a fairer comparison to a more robust baseline. The tuning (implemented through a grid search) and training of all models (*n* embeddings x *m* cognitive data sources) are fully automatized within the command `run`.

[6] `https://github.com/facebookincubator/python-nubia`

Thereafter, the user can either use the saved results as they are (i.e., mean squared errors for each word in the vocabulary), or they can run the significance testing (command: `significance`), which consists of a Wilcoxon signed-rank test for each hypothesis (i.e., for each embedding / cognitive-data evaluation combination), applying the Bonferroni correction for the multiple hypotheses problem, as described by Dror et al. (2018). Finally, the automatized generation of the report also includes significance testing by default and compares the results to the baseline of random embeddings before aggregating them. The dynamic HTML or PDF reports include all detailed results for the individual combinations, as well as aggregated over the modalities (see Figure 6.9a). Table 6.9 presents the most important commands provided in the CogniVal CLI. A full tutorial is provided in the GitHub repository.[7]

6.6.2 Use Cases

The target audience for the CogniVal CLI are NLP and machine learning practitioners and researchers developing word embeddings and in need of an evaluation benchmark. We describe the following two possible use case scenarios.

Scenario 1: Custom Word Embeddings & Cognitive Data Sources

One of the most relevant features of the CogniVal CLI is the possibility to upload custom word representations from any language model. Any type of word embedding can be imported into the system as text or binary files, and can then be evaluated against the available cognitive data sources and compared to the other embeddings included in CogniVal by default. Moreover, the automated versioning and reporting supports the development process of new embeddings by readily generating plots over the course of time to show whether the performance of the embeddings in development is improving or deteriorating across multiple runs.

In addition, custom cognitive data sources can also be imported into the CogniVal interface. This feature allows the user to add more cognitive language processing data as more of these datasets become available (Alday, 2019). Through these features, CogniVal becomes a generic

[7] `https://github.com/DS3Lab/cognival-cli/blob/master/cognival_tutorial.pdf`

Interaction step	Example command(s)

1. Import data sources	`$ import cognitive-sources source=yourCustomSource`
	`$ import embeddings youCustomEmbeddings.zip`
	`$ import embeddings glove.6B.50`
	`$ import random-baselines glove.6B.50 num-baselines=10`

2. Define evaluation combinations	`$ config experiment cognitive-sources=[eeg_zuco] embeddings=[glove.6B.50]` `$ config experiment cognitive-sources=[eye tracking_geco] embeddings=[fasttext]`
3. Fit regression models	`$ run embeddings=[glove.6B.50,glove.6B.100] cognitive-sources=[eye tracking_geco] cognitive-features=[WORD_FIXATION_COUNT]`
4. Perform significance testing	`$ significance run_id=0 modalities=[eye tracking, eeg] alpha=0.01 test=Wilcoxon`
5. Generate report	`$ report open-html=True`

Table 6.9: Main steps and example commands for using the CogniVal CLI for cognitive word embedding evaluation.

framework for cognitive word embedding evaluation and drastically increases the number of possible applications in any language.

Scenario 2: Complementary Benchmark & Evaluation Over Time

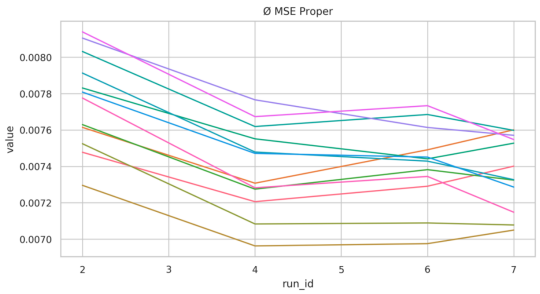
The CogniVal CLI also permits to use the cognitive word embedding evaluation as a complementary evaluation and a benchmark for embedding selection. If the user wants to compare the results achieved by their embeddings on CogniVal to other intrinsic or extrinsic results achieved with the same embeddings, the CLI allows uploading external results into the result directory and runs the significance testing and aggregation report on all available results. Hence, CogniVal can easily be extended to include external results and can be leveraged as a tool for embedding selection. Furthermore, CogniVal can be used during the development of language models. If one is comparing a certain checkpoint of a language model to extrinsic results on a downstream task, the cognitive evaluation can help to ensure that the word representations are not overfitting on the downstream task and that they still maintain cognitive plausibility. To this end, the automatic report generated in the CogniVal CLI also includes plots to show changes over various

Results (Aggregated per Embedding and Modality)

Eye-Tracking

Word embedding	Ø MSE Baseline	Ø MSE Proper	Significance
bert1	0.13905	0.00740	42/42
bert2	0.13974	0.00760	42/42
bert3	0.14057	0.00705	42/42
bert4	0.14013	0.00708	42/42
bert5	0.13949	0.00732	42/42
bert6	0.14030	0.00753	42/42
bert7	0.13835	0.00760	42/42
bert8	0.13772	0.00733	42/42
bert9	0.14116	0.00729	42/42
bert10	0.13918	0.00757	42/42
bert11	0.13794	0.00755	42/42
bert12	0.13936	0.00715	42/42

(a) Automatically generated result table.



(b) Plot over time: When adding more eye tracking features in each run, the aggregated results become more precise. Each colored line represents a different BERT layer.

Figure 6.9: Snippets from an automatically generated result report in the CogniVal CLI.

runs, which can be very useful during the development of new languae models or fine-tuning of existing pre-trained language models (see Figure 6.9b).

6.6.3 Example Application: Comparison of BERT Layers

As an exemplary use case scenario for the CogniVal CLI, we analyze the performance of different layers of BERT pre-trained contextual word representations on all available cognitive data sources of eye tracking and EEG. Transformer-based language models such as BERT are widely used in state-of-the-art NLP, but their inner workings are still largely unknown (Rogers et al., 2021). We extract word-level BERT embeddings (Devlin et al., 2019) for all words where there is cognitive data available. Using the `bert-as-service` package[8] we extract the hidden states of all 12 layers of the `BERT-base-uncased` model of 768 dimensions. Subsequently, using the new CogniVal functionality to load custom embeddings, we import the BERT states of each layer easily into the CogniVal interface. We set the configuration to run all experiments against the 10-fold random baseline, with the following parameters for training: 3 fold cross-validation,

[8] `https://github.com/hanxiao/bert-as-service`

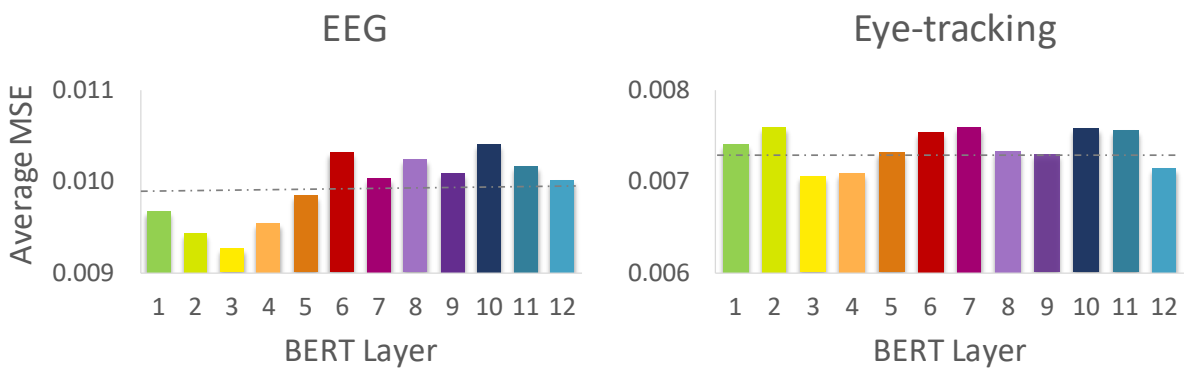


Figure 6.10: CogniVal results on evaluating the 12 layers of a BERT model against all available EEG and eye tracking data sources.

a hidden layer of 200 dimensions, 20% validation split, batch size of 128, and ReLu activation functions.

The results of this example application are presented in Figure 6.10. In addition, Figure 6.9a shows the numerical summary of the results of BERT embeddings predicting eye tracking features as it is presented in the automatically generated report, including the results of the random baselines and the results of the significance testing. While all hypotheses tested on the BERT layers proved to be statistically significant against the random baseline (4 EEG hypotheses – one for each dataset, and 42 eye tracking hypotheses – one for each feature), there are visible differences in performance between the layers. Surprisingly, for both EEG and eye tracking, layer 3 performs best. The results also show how the last layer performs very closely to the average of all layers (dashed line). This finding reflects the original performance of the layers of the BERT base model on downstream NLP tasks (Devlin et al., 2019). It is also in line with Toneva and Wehbe (2019), who find that the lower layers perform best at predicting neural activation for short context ranges. Lin et al. (2019) show that the lower layers have the most linear word order information, which is likewise reflected in our results. This application scenario shows how Cognival can be used to explore and support findings concerning the interpretability of language models.

6.7 Summary

We presented CogniVal, the first multi-modal large-scale cognitive word embedding evaluation framework. The vectorized word representations are evaluated by using them to predict eye tracking or brain activity data recorded while participants were understanding natural language. We find that the results of eye tracking, EEG and fMRI data are strongly correlated not only across these modalities but even between datasets within the same modality. Intriguingly, we also find a correlation between our cognitive evaluation and two extrinsic NLP tasks, which might indicate that CogniVal could also be used for predicting downstream performance and hence, choosing the best embeddings for specific tasks.

We plan to expand the collection of cognitive data sources as more of them become available, including data from other languages such as the Narrative Brain Dataset (Dutch, fMRI, Lopopolo et al. (2018)) or the Russian Sentence Corpus (eye tracking, Laurinavichyute et al. (2019)). Thanks to naturalistic recording of longer text spans, CogniVal can also be extended to evaluate sentence embeddings or even paragraph embeddings. CogniVal can become even more effective by combining the results with other intrinsic or extrinsic embedding evaluation frameworks (Nayak et al., 2016; Rogers et al., 2018) and building on the multiple hypotheses testing.

In Section 6.6, we presented the command line interface for CogniVal. The CogniVal CLI builds upon the work presented in Section 6.4 and extends it with various new features, especially the ability to evaluate custom embeddings against custom cognitive data sources. We described the functionalities of the tool as well as various use cases and an application scenario. CogniVal is still under active research and will be extended to additionally support the evaluation of sentence embeddings and further languages. Moreover, the CLI aims at improving the accessibility and usability of cognitive embedding evaluation for NLP practitioners.

Chapter 7

Conclusions and Future Work

Human language processing signals can be leveraged to improve and evaluate machine learning models of natural language understanding. An important aspect to be discussed is the ethical implication of this type of work. We do so in Section 7.1 before summarizing the contributions of this thesis (Section 7.2) and concluding with possible future research avenues (Section 7.3).

7.1 Ethical Considerations

To conclude this thesis, we address some of the ethical considerations that arise when working with human language processing signals for NLP. As researchers in this area, we mostly make use of existing datasets that have been collected by psychology researchers. Nevertheless, the following ethical aspects should be taken into account.

First, we want to highlight the necessity of considering the high-level consequences of our work. It becomes increasingly relevant to examine the implications of the interaction between humans and machines, between what can be recorded from a human brain and what can be extracted from those signals. What is the potential of the derived results? What is the objective of the final application? What is the impact on people and society? Suster et al. (2017) describe this aspect as the dual use of data: Applications leveraging cognitive cues for improving NLP (and many other machine learning applications) have the potential to be applied in both beneficial and harmful ways.

Second, it is essential to remember the responsibility towards research subjects and towards protecting the individual (Suster et al., 2017). All collected data comes from humans willing to share their brain activity for research. Hence, the participants as well as their data should be treated respectfully, even if as NLP practitioners we are leveraging the provided data and not recording it ourselves. Although the data is anonymized after recording, we should refrain from drawing inferences from our models back to single participants.

Finally, the origins of the data and any biases within them should be considered. Most psychological studies are based on Western, educated, industrialized, rich, and democratic research participants (so-called *WEIRD*, Henrich et al. (2010)). By assuming that human nature is so universal that findings on this group would translate to all other demographics, this has led to a heavily biased collection of psychological data. The potential consequences of exclusion or demographic misrepresentation should not be ignored (Hovy and Spruit, 2016). One step further, Caliskan et al. (2017) showed that text corpora contain recoverable and accurate imprints of our historic biases. These biases can be extracted from the text, and are also reflected in eye movements and brain activity recordings (Wu et al., 2012; Herlitz and Lovén, 2013; Fabi and Leuthold, 2018). Thus, it is essential to remember that with extensive reuse of the same corpora, these biases – participant sampling as well as experimental biases – are propagated to many experiments, and researchers should be careful in the interpretation of the results.

7.2 Summary

Computational models for natural language understanding have made tremendous progress over the past few years. However, these ML models are imperfect and lack intricate capabilities that humans access automatically when processing speech or reading text. Human language processing data can be leveraged to increase the performance of models and to pursue explanatory research for a better understanding of the differences between human and machine language processing. In this dissertation, we focused on approaching this goal from an interdisciplinary stance and can be divided into three threads: (1) compiling and sharing datasets of naturalistic cognitive language processing tailored to NLP, (2) leveraging human language processing data to improve natural language processing methods, and (3) evaluating the cognitive plausibility of word representations used for NLP.

When we read, our brain processes language and generates cognitive processing data such as

gaze patterns and brain activity. These signals can be recorded with various methods such as eye tracking and brain activity. As described in Chapter 2, we led the efforts of compiling the *Zurich Cognitive Language Processing Corpus* (ZuCo), a dataset of simultaneous electroencephalography (EEG) and eye tracking recordings from subjects reading natural sentences. ZuCo 1.0 and ZuCo 2.0 in total include data of 30 healthy adult native English speakers, each reading natural English text for 4–6 hours covering between 700 and 1100 sentences. This dataset represents a valuable resource for cognitively-inspired natural language processing, and the co-registration of eye movements and brain activity allows for linguistic analyses on multiple levels, from lexical processing to full sentence processing.

This data collection effort established the basis of this thesis. It allowed us to pursue research questions such as: Can eye movement signals improve higher-level semantic tasks such as extracting information from text? From the vast body of psycholinguistic research, it is known that word frequency and word familiarity influence how long readers look at a word. Therefore, we conducted a study of improving named entity recognition with eye tracking signals (see Chapter 3). We leveraged eye movement features from corpora with recorded gaze information to augment a state-of-the-art neural model for named entity recognition with gaze embeddings. Moreover, we show how gaze features, generalized to the word type level as well as learning to predict cognitive features, eliminate the need for recorded eye tracking data at test time.

Similar work using eye tracking features to improve other NLP tasks such as part-of-speech tagging (Barrett et al., 2016) or sentiment analysis (Mishra et al., 2017a) has been carried out. In Section 3.3, we analyzed whether using such human features can show consistent improvement across tasks and data sources. We presented an extensive investigation of the benefits and limitations of using cognitive processing data for NLP, as well as a review of best practices in Section 3.4. As another method of improving NLP methods with human data, we have used eye tracking and EEG data as an inductive bias in a multi-task learning architecture (Barrett et al., 2018a; Muttenthaler et al., 2020). The attention mechanism is trained on human data and shows improvements over standard machine attention in neural networks. However, this work is outside the scope of this thesis.

Not much work had been conducted on brain activity data for NLP since the ample range of cognitive processes and low signal-to-noise ratio presents many challenges when leveraging brain activity data for NLP. Hence, signals such as EEG brainwaves are still largely unexplored in this context. In a large-scale study presented in Chapter 4, we systematically analyzed the potential of electrical brain activity data for improving natural language understanding tasks, which special focus on feature extraction, word embedding types, task complexity and training data size.

The multi-modal machine learning models we implemented leveraging EEG data achieved consistent improvements across different embedding types, but the magnitude decreases for more difficult tasks. Hence, human signals again showed great potential, but also make evident the need to develop better machine learning algorithms for effectively combining multiple input modalities.

In Chapter 5, we investigated to what extent state-of-the-art pre-trained language models entail human sentence processing patterns in the form of eye movement data. We compared and analyzed the performance of language-specific and multilingual pre-trained transformer language models to predict behavioral measures that reflect human sentence processing on Dutch, English, German, and Russian texts. We find that BERT and XLM language models can accurately predict a range of gaze features and that they successfully reflect properties found in human reading.

Finally, we used cognitive language processing signals to evaluate word representations, the cornerstones of state-of-the-art natural language processing (see Chapter 6). It has been noted that current intrinsic evaluation methods do not tell us anything about the cognitive plausibility of word embeddings and language models (Gladkova and Drozd, 2016). Therefore, we developed *CogniVal*, the first openly available framework for evaluating English word embeddings based on cognitive lexical semantics. Specifically, embeddings are evaluated by their performance at predicting eye tracking, EEG and fMRI data recorded during language comprehension. This framework is easily extensible and available to include other intrinsic and extrinsic evaluation methods, which is essential for achieving a global evaluation metric. We find strong correlations in the results between cognitive datasets, across recording modalities and to their performance on extrinsic NLP tasks. However, further research is required to provide a global intrinsic evaluation framework based on human language processing data, which is able to interpret and relate the linguistic structures in both human and machine learning models (Manning et al., 2020).

During past and current projects, we have contributed to this field and gained insights into the interplay between machine learning for NLP and human language processing. Some of the remaining open challenges reflected in this line of research are the effective combination of multiple behavioral modalities, the extension to a wider range of languages, and interpretable machine learning solutions. Based on the current state of this research field and the work presented in this thesis, we have identified several avenues for future research that we describe below. Moreover, we shortly discuss the ethical considerations to be taken into account in this line of research.

7.3 Directions for Future Research

Current state-of-the-art machine learning algorithms for language understanding still lack certain skills that human are able to perform automatically and effortlessly. These skills include generalization, multi-modal learning, and learning complex tasks. Our experience of the world is multi-modal. We learn to speak by listening and observing. We communicate via various channels including speech, gestures, and facial expressions. On the one hand, humans are able to learn new things from very few examples, especially in language acquisition. On the other hand, machine learning methods still struggle to generalize and to learn from multiple modalities simultaneously. So far, multi-modal learning in natural language processing mostly takes into account additional input types such as images or audio. However, as indicated by the work in this thesis, it is important and promising to bring human signals into the picture.

This research is highly interdisciplinary and falls into the intersection of natural language processing, machine learning, and cognitive science. Based on our previous research combining human language processing data such as eye tracking and brain activity with natural language understanding methods, we found that leveraging cognitive data for NLP is very promising and shows great potential for multi-modal machine learning as well as for interpretation approaches. However, there are still open challenges: (1) Human data is extremely noisy. (2) Compared to multi-modal approaches leveraging additional modalities such as images, which are based on learning from co-occurrence information, human data is rich in cognitive processing information and linguistic structure. (3) The fact that there are known neuropsychological differences in processing different languages, but most NLP research still focuses only on English, shows the need for cross-lingual and multilingual approaches.

Ultimately, the goal of this line of research is to bridge human intelligence with machine intelligence to build a general, interpretable framework for multi-modal NLP using a diverse range of cognitive signals. We believe that through human-grounded learning we can build truly generalizable models of natural language processing. We have shown how human language processing signals can be leveraged to increase the performance of models and to pursue explanatory research for a better understanding of the differences between human and machine language processing. We hope that the insights of this work will help benefit the field of cognitively-inspired natural language processing, and that we could make a small contribution to further our understanding of human and machine language understanding.

Chapter A

Appendices

A.1 Evaluation Metrics

In this section, we shortly define the evaluation metrics used in this thesis.

We use *precision*, *recall*, and F_1-*score* to evaluate the classification models. For classification tasks, the terms *true positives* (TP), *true negatives* (TN), *false positives* (FP), and *false negatives* (FN) compare the predictions of the classifier during testing to the ground truth. The terms *positive* and *negative* refer to the classifier's prediction, and the terms *true* and *false* refer to whether that prediction corresponds to the ground truth label. Precision is the fraction of relevant instances among the retrieved instances and is defined as

$$Precision = \frac{TP}{TP + FP} \tag{A.1}$$

Recall is the fraction of the relevant instances that are successfully retrieved:

$$Recall = \frac{TP}{TP + FN} \tag{A.2}$$

The F_1-score is the harmonic mean combining precision and recall:

$$F_1\ score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{A.3}$$

For the evaluation of the regression models, we use the mean absolute error (MAE) and the mean squared error (MSE). The MAE of a model with respect to a ground truth test set is the

mean of the absolute values of the individual prediction errors over all instances in the test set. Each prediction error is the difference between the true value x_i and the predicted value y_i for the instance:

$$MAE = \frac{\sum_{i=1}^{n} \mid y_i - x_i \mid}{n} \tag{A.4}$$

where, n is the total number of instances. Analogously, the MSE measures the average of the squares of the errors, i.e., the average squared difference between the predicted values and the true values:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2 \tag{A.5}$$

A.2 ZuCo: Technical Data Validation

The aim of the technical validation of the data included in the Zurich Cognitive Language Processing Corpus (Chapter 2) is to guarantee good recording quality and to replicate the findings of previous studies investigating co-registration of EEG and eye movement data during natural reading tasks (e.g., Dimigen et al. (2011)). We also compare the results between ZuCo 1.0 and Zuco 2.0, which allows a more direct comparison due to the analogous recording procedure.

Eye Tracking

Omission rates & skipping proportions: The eye tracking data were evaluated by analyzing the fixations made by each subject. On the one hand, we analyze the fixations on sentence level using the omission rate. The omission rate is defined as the percentage of words that are not fixated in a sentence. Figure A.1 (left) shows the omission rates per task for each subject in ZuCo 1.0 and A.2 (left) for ZuCo 2.0. Clearly, the participants made fewer fixations during the task-specific reading, which led to faster reading speed. On the other hand, we analyze the skipping proportion on the word level. The skipping proportion is the rate of words being skipped (i.e., not being fixated). Figure A.1 (right) presents the skipping proportion for all tasks

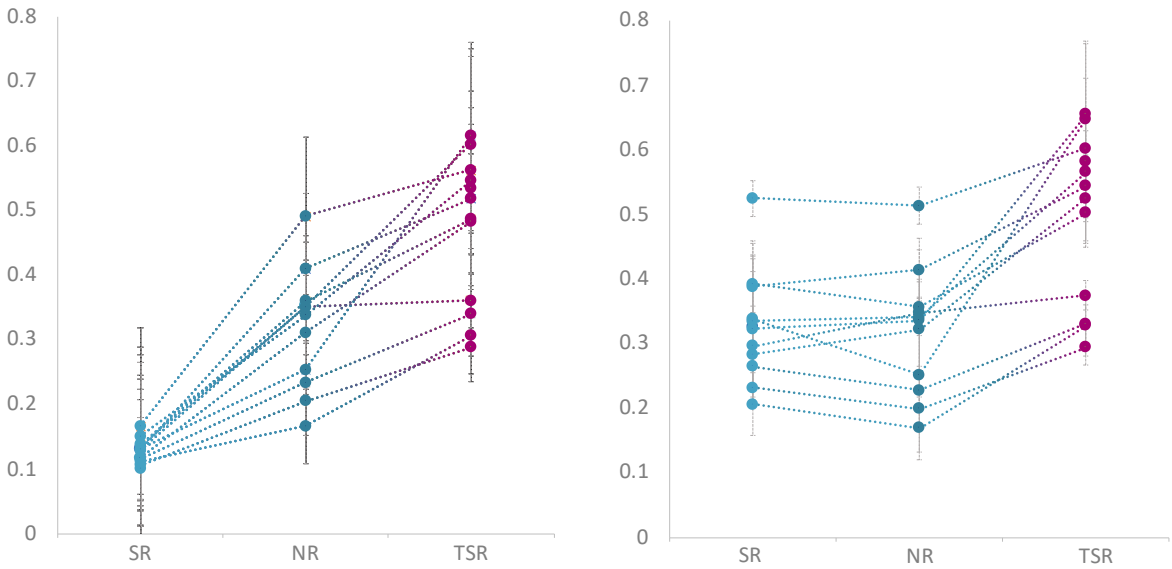


Figure A.1: Omission rates and skipping proportions (means and standard errors) for all tasks and subjects (means and standard error) in ZuCo 1.0. (left) The omission rates for each task and for each subject, where the y-axis shows the proportion of words being skipped in a sentence (0–1). (right) The skipping proportion (y-axis) for each task and for each subject.

and all subjects in ZuCo 1.0, and Figure A.2 (right) for ZuCo 2.0, respectively. As expected, this also increases in the task-specific reading. The values of the reported metrics are in accordance with values reported in other eye tracking reading studies (Cop et al., 2017; Rayner, 1998). As mentioned above, both the omission rate and the skipping proportion are significantly higher in the task-specific reading paradigm. Because the readers were searching for a particular relation in this annotation task, this does not necessarily require reading all words in a sentence until the end.

Moreover, we present the effect of the word length on the skipping proportion for both ZuCo 1.0 and ZuCO 2.0, as it has been presented previously by Cop et al. (2017). Figure A.3 shows that the probability of a word being skipped decreases for longer words, but short words are frequently skipped consistently in all tasks.

Reading times: The reading times of the recorded data were also validated. As described previously, we extracted the features described in the GECO corpus (Cop et al., 2017): first fixation duration (FFD), single fixation duration (SFD), gaze duration (GD), total reading time

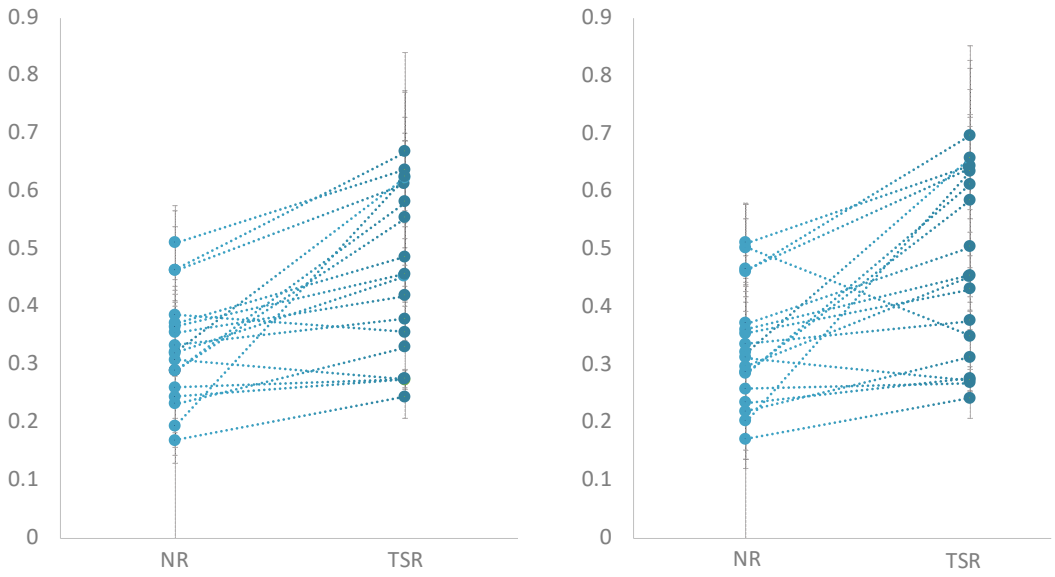


Figure A.2: Omission rates and skipping proportions (means and standard errors) for all tasks and subjects (means and standard error) in ZuCo 2.0. (left) The omission rates for each task and for each subject, where the y-axis shows the proportion of words being skipped in a sentence (0–1). (right) The skipping proportion (y-axis) for each task and for each subject.

(TRT) and go-past time (GPT). This is visualized in Figure A.4 for ZuCo 1.0 and in Figure A.5 for ZuCo 2.0, which shows the mean and distribution of the reading times for each feature, separated by task. For reference, Table A.1 presents the exact values. For the calculation of these reading times, only fixations $> 100ms$ were considered. The presented ranges are in line with those presented by Cop et al. (2017), for both ZuCo 1.0 and 2.0.

Although the reading material is from the same source and of the same length range (see Figure A.11 (left)), in the first task (NR) passive reading was recorded, while in the second task (TSR) the subjects had to annotate a specific relation type in each sentence. Thus, the task-specific annotation reading led to shorter passes because the goal was merely to recognize a relation in the text, but not necessarily to process every word in each sentence. This distinct reading behavior is shown in Figure A.10 for ZuCo 2.0, where fixations occur until the end of the sentence during normal reading, while during task-specific reading the fixations stop after the decisive words to detect a given relation type.

Feature	Normal Reading (Sentiment)			Normal Reading (Wikipedia)			Task-specific Reading (Wikipedia		
	M	SD	Range	M	SD	Range	M	SD	Range
FFD	229	49	100-984	223	61	100-1542	219	57	100-988
SFD	235	59	100-984	227	70	100-1542	222	62	100-916
GD	263	84	100-1287	267	110	100-1880	243	80	100-1206
TRT	359	168	100-1839	371	199	100-3313	306	146	100-2100
GPT	445	363	100-5291	452	419	100-6935	399	383	100-13128

Table A.1: Mean (M), standard deviations (SD), and ranges of the reading time measures per feature for each task in milliseconds.

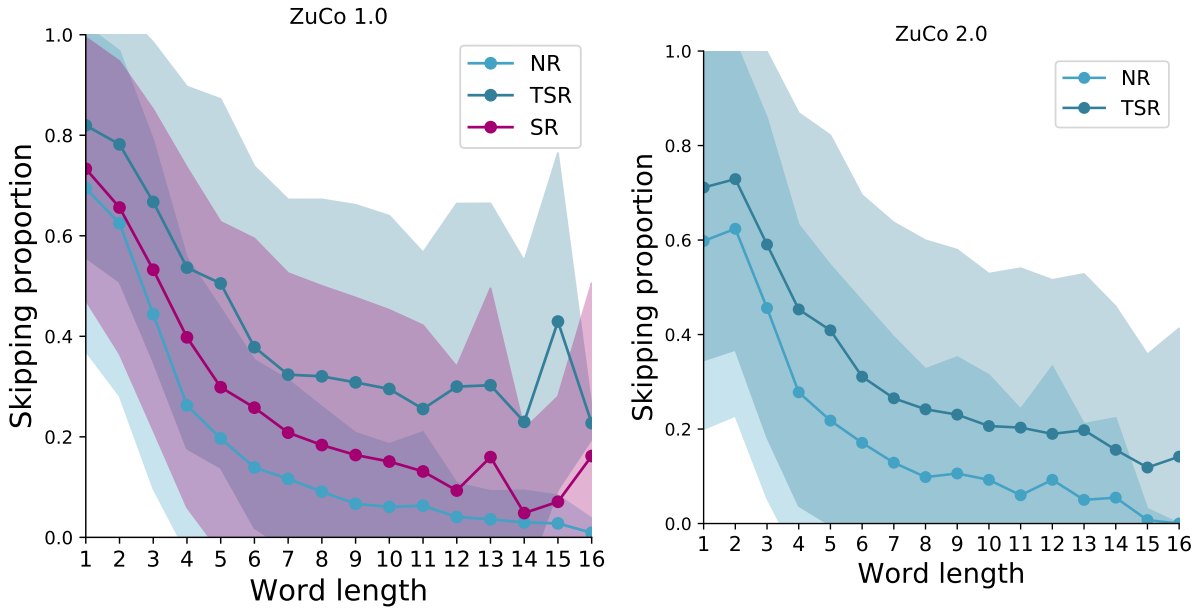


Figure A.3: Effect of word length on the skipping proportion per task (mean and standard deviation), x-axis = word length, y-axis = mean skipping proportion.

Electroencephalography (EEG)

The aim of the technical validation of the EEG data was to replicate the findings of previous studies investigating co-registration of EEG and eye movement data during free reading tasks (Dimigen et al., 2011; Rayner, 1998). As shown in Figure 2.5, the difference between normal

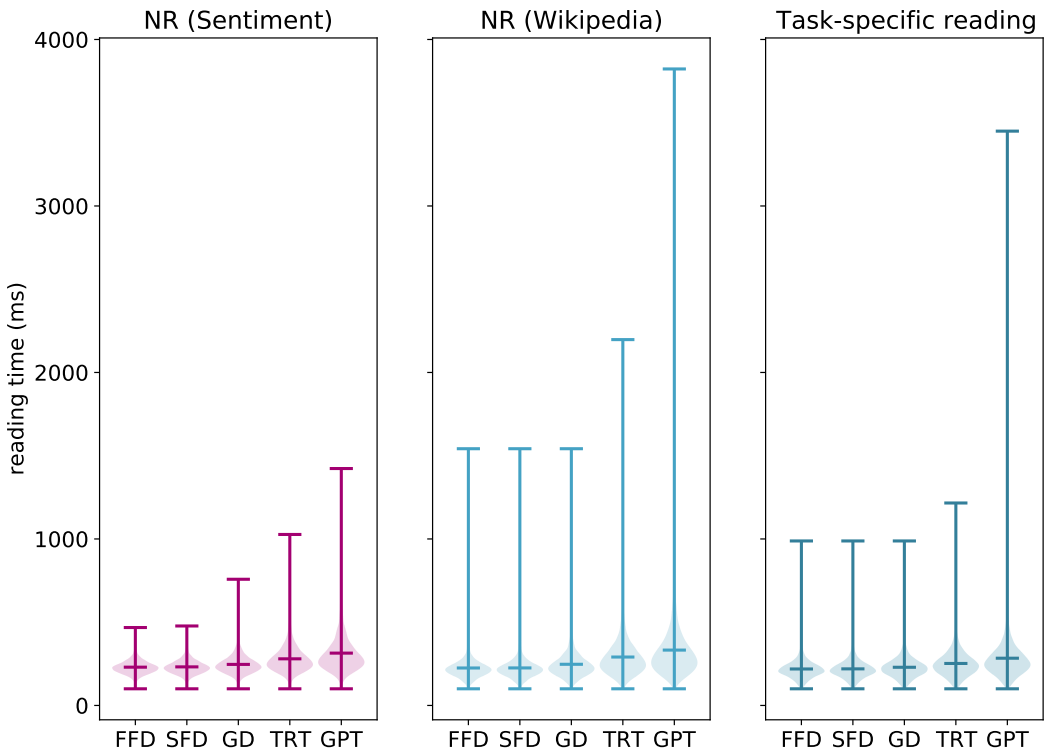


Figure A.4: Violin plot showing means, distributions, and ranges of the reading time measures per word for each task in ZuCo 1.0 and each eye tracking feature (x-axis) in milliseconds.

reading and task-specific annotation reading is also evident in the brain activity data.

Preprocessing of the EEG data was analogous to the feature extraction; however, after preprocessing, a single bandpass filter was applied in addition (0.5–30 Hz). Afterwards, EEG data were re-referenced to the average reference and segmented to the onsets of the fixations. Analogous to Dimigen et al. (2011), a time-window of 1600 ms was chosen (600 ms before onset of the fixation to 1000 ms after). Furthermore, the pool of fixations was limited to first-pass reading fixations, which yielded in 154,173 trials in total.

Fixation-related potentials: As a first validation step, fixation-related potentials (FRPs) were extracted. Therefore, the pool of fixations described above was divided into the three task conditions (Task 1, n=56,743; Task 2, n=48,723; Task 3, n=48,707) and trials were averaged within each condition. Data were baseline corrected using a 100 ms time-window before fixation on-

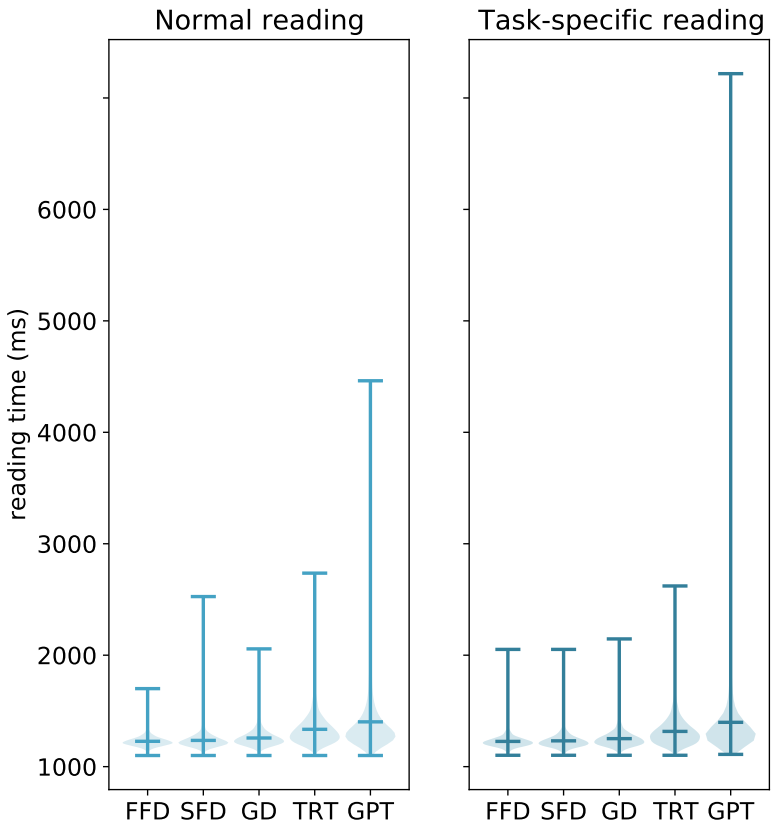


Figure A.5: Violin plots showing means, distributions, and ranges of the reading time measures per word for each task in ZuCo 2.0 and each eye tracking feature (x-axis) in milliseconds.

set. Figure A.6 shows the time-series of the resulting FRPs for two electrodes (PO8 and Cz) as well as the topographies of the voltage scalp distributions at selected timepoints for ZuCo 1.0, and Figure A.8 for ZuCo 2.0. Due to the different EEG system in the study of Dimigen et al. (2011), electrode locations did not match perfectly (Henderson et al., 2013). To compare the results, we used similar electrodes. The time-points were chosen according to Dimigen et al. (2011), namely, the visually evoked lambda response of the previous fixation (1), the mygoenic spike potential at saccade onset (2), the lambda response of the current fixation (3), the N170 component (4) and the N400 component, which is overlapped by the lambda response of the succeeding fixation (5).

All results are in line with the findings of Dimigen et al. (2011). The five ERP components (for

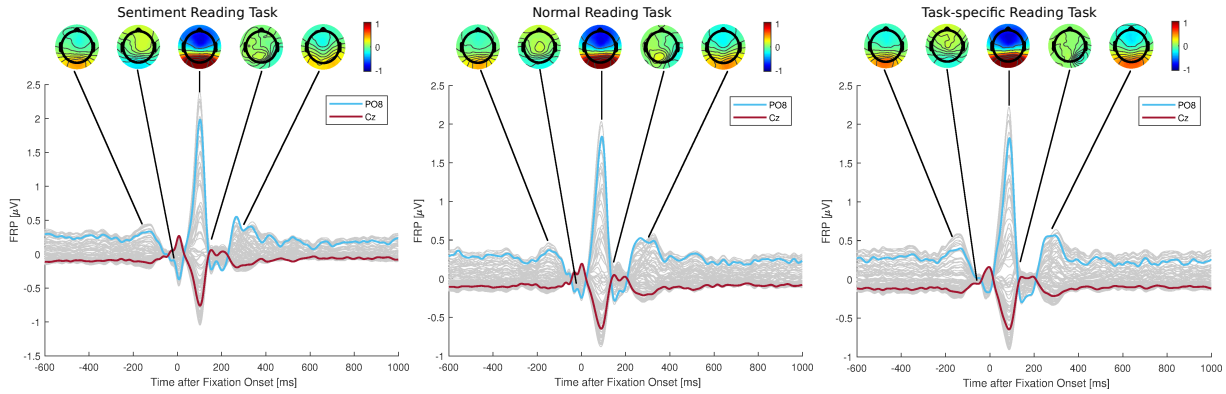


Figure A.6: ZuCo 1.0: FRPs during the different task conditions with selected scalp level potential distributions. Topographies show amplitudes in microvolt, coded as color.

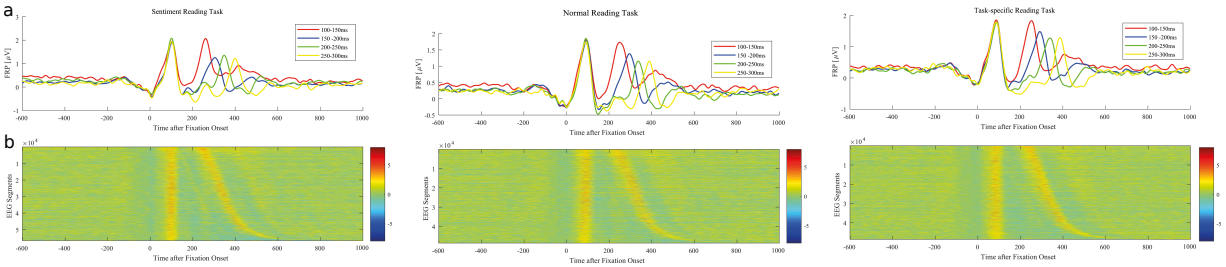


Figure A.7: ZuCo 1.0: Clustered EEG segments. (a) FRPs of electrode Cz, clustered by duration of the fixation. (b) Each horizontal line represents the mean of the current and 50 adjacent EEG epochs, segmented on fixation onset. Segments are ordered by fixation duration (top: shortest fixation, bottom: longest fixation). Color represents the amplitude of the signal in microvolt.

which the scalp topographies are plotted) are highly similar to this study in the time-course of the chosen electrodes. This also applies for the scalp level topographies.

Fixation duration effect on ERPs: Previous studies were able to show an effect of fixation duration on the resulting FRPs (Dimigen et al., 2011; Henderson et al., 2013). We followed two approaches to demonstrate this dependency in the current dataset. We first followed the procedure of Dimigen et al. (2011), therefore all single-trial FRPs were ordered by fixation duration. As a next step, a vertical sliding time-window was used to smooth the data; that is, for each EEG segment, the average of 50 adjacent trials was calculated. Figure A.7 (a) and Figure A.9 (a) show the resulting plots per task condition for ZuCo 1.0 and ZuCo 2.0, respectively. In line with the previous study by Dimigen et al. (2011) a first positivation P1 can be identified at 100 ms post-fixation onset. A second positive peak P2 is located dependent on the duration of

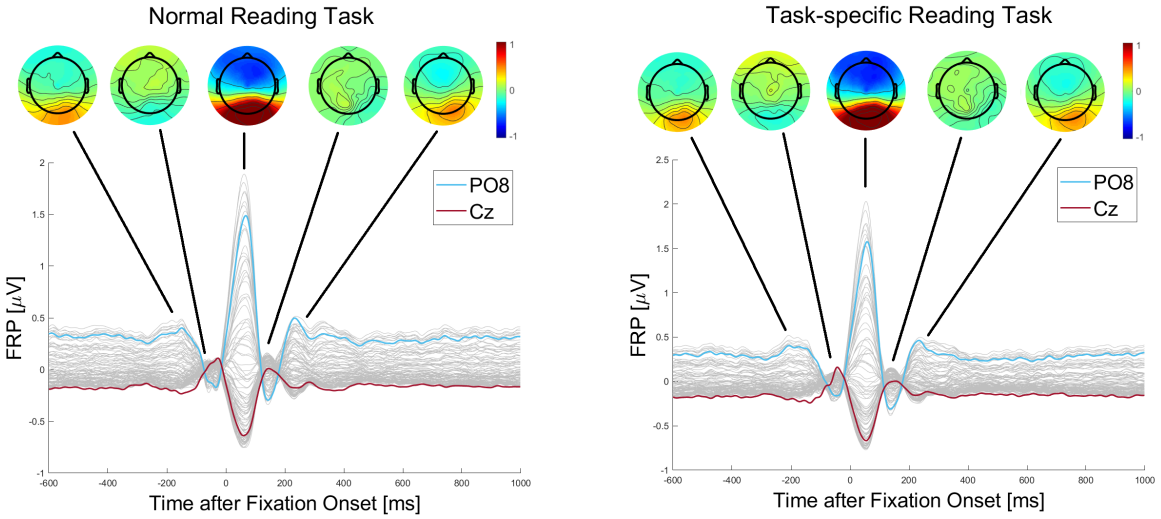


Figure A.8: ZuCO 2: Fixation-related potentials (FRPs) during both task conditions with selected scalp level potential distributions. Topographies show color-coded amplitudes in microvolt.

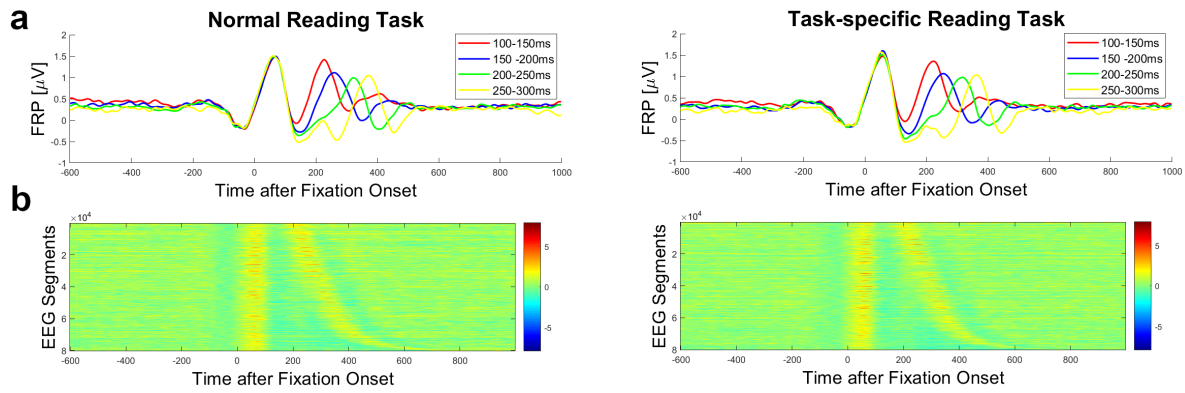


Figure A.9: ZuCO 2: Clustered EEG segments. (a) FRPs of electrode *Cz*, clustered by duration of the fixation. (b) Each horizontal line represents the mean of the current and 50 adjacent EEG epochs, segmented on fixation onset. Segments are ordered by fixation duration (top: shortest fixation, bottom: longest fixation). Color represents the amplitude of the signal in microvolt.

the fixation, which can be explained by the time-jittered succeeding fixation.

The second approach is based on previous related work (Luu and Ferree, 2005), in which single trial EEG segments are clustered by the duration of the current fixation. Here, four clusters were chosen (100–150 ms, 150–200 ms, 200–250 ms, 250–300 ms). Data within each cluster were averaged to get four different FRPs, depending on the fixation duration. The results can be seen in Figure A.7 (b) for ZuCo 1.0, and Figure A.9 (b) for ZuCo 2.0. While P1 is located around

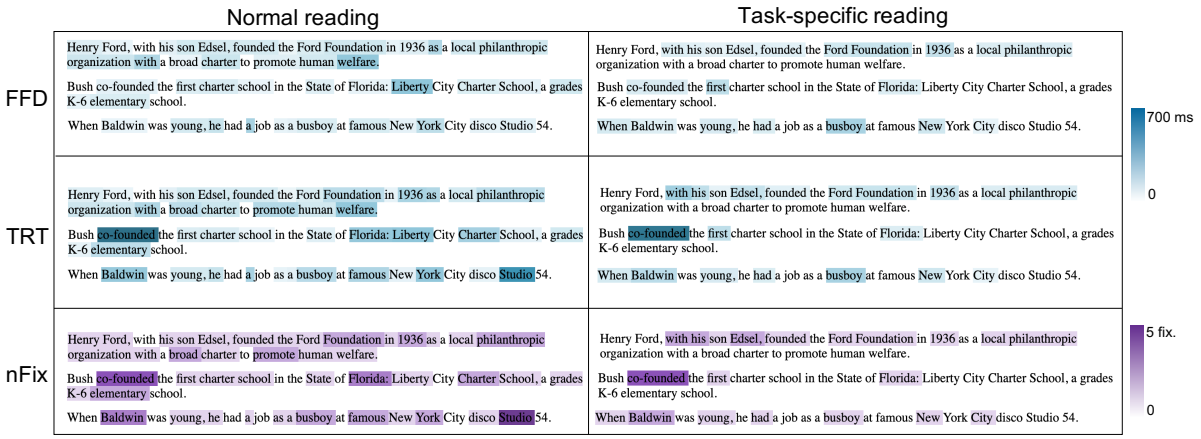


Figure A.10: Fixation heatmaps for three sentences in ZuCo 2.0 containing the relation *founder*, showing a comparison of the eye tracking features between normal reading and task-specific annotation reading for a single subject (first fixation duration (FFD), total reading time (TRT), number of fixations (nFix).

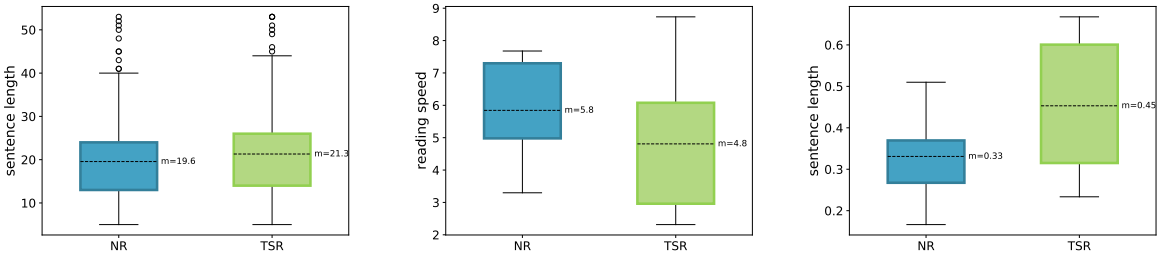


Figure A.11: Sentence length (words per sentence), reading speed (seconds per sentence) and omission rate (percentage of words not fixated) comparison between normal reading (NR) and task-specific reading (TSR) in ZuCo 2.0.

100 ms post-fixation onset, the P2 component also moves as a function of the fixation duration, which is consistent with the findings reported by Henderson et al. (2013).

A.3 Additional Results for Eye Tracking Prediction

In this section we present the implementation details as well as additional results for the experiments on the prediction of human reading behavior presented in Chapter 5.

Eye Tracking Data

Figure A.12 shows the word length effect on the eye tracking data. i.e., the fact that longer words are more likely to be fixated. This effect is observable across all languages.

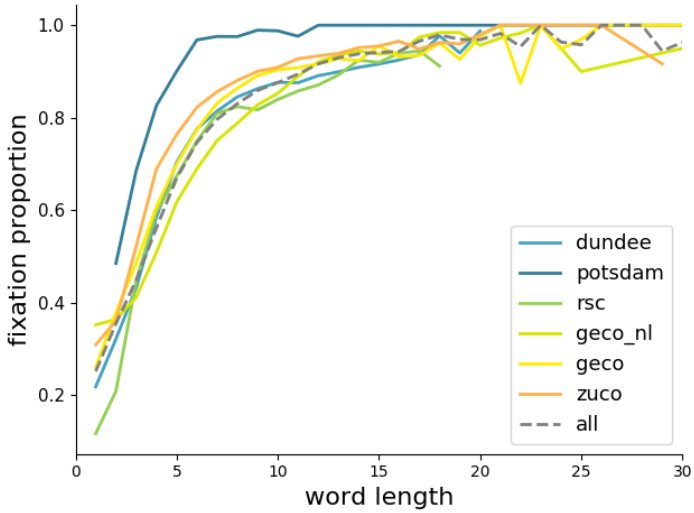


Figure A.12: Word length effect on all datasets in all four languages.

Figure A.13 shows the mean fixation duration (MFD) for adjectives (ADJ), nouns (N), verbs (V) and adverbs (ADV) for all six datasets. We use Spacy[1] to do part-of-speech tagging for our analyses. For Russian we load an externally trained model[2], for Dutch, English and German we use the provided pre-trained models.

Readability Scores

We use the Flesch Reading Ease (FRE) score (Flesch, 1948) to define the readability of the English text in the eye tracking corpora. This score indicates how difficult a text passage is to understand based on the average number of words in a sentences and the average number of syllables in a word:

[1] `spaCy.io`
[2] `https://github.com/buriy/spacy-ru`

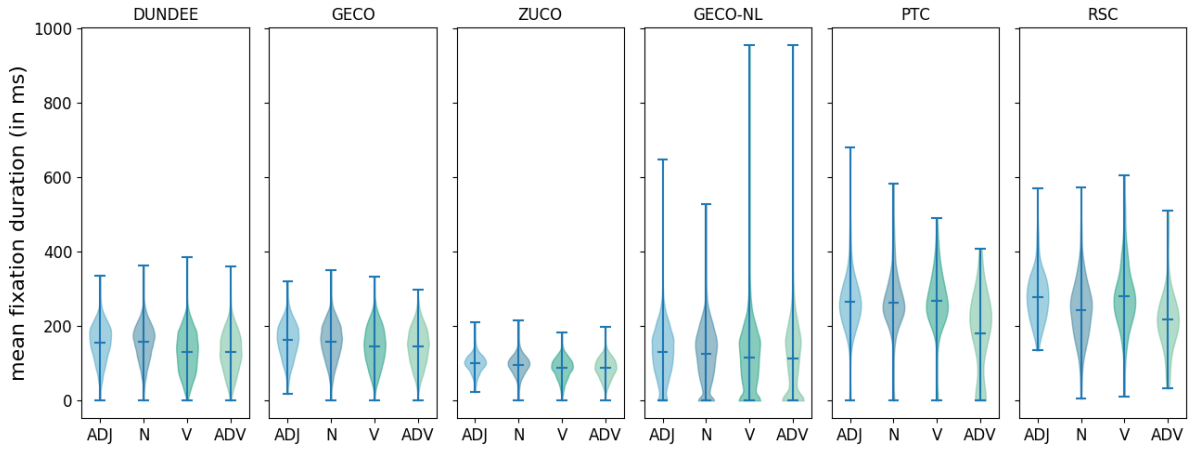


Figure A.13: Distribution of mean fixation duration (MFD) values for the most common parts of speech (adjectives (ADJ), nouns (N), verbs (V) and adverbs (ADV)) across all six datasets.

$$FRE = x - y\left(\frac{words}{sentences}\right) - z\left(\frac{syllables}{words}\right) \tag{A.6}$$

Since the FRE score relies on the language-specific weighting factors x, y and z, we apply the Flesch Douma adaptation for Dutch (Douma, 1960), the adaptation by Amstad (1978) for German, and the adaptation by Oborneva (2006) for Russian. We used the implementation by `https://github.com/wimmuskee/readability-score` for English and Dutch, and added the implementations for German and Russian based on the adaptations mentioned above.

Model	Batch size
BERT-En, BERT-Nl, BERT-Multi	16
BERT-De, BERT-Ru, XLM-EnDe, XLM-17, XLM-100	8
XLM-En	2

Table A.2: Batch sizes used for each of the language models.

Implementation Details

Tokenizer When using BERT for token classification or regression, a pressing implementation issue is presented by the *WordPiece* tokenizer employed by the model (Wu et al., 2016). This tokenizer handles unknown tokens by recursively splitting every word until all subtokens belong to its vocabulary. For example, the name of the Greek mythological hero "Philammon" is tokenized into three subtokens "['phil', '##am', '##mon']". In this case, our BERT model for token regression would produce an eight-dimensional output for all three subtokens, and we had the choice as to what to do in order to compute the loss, having only one target for the full word "Philammon". We chose to compute the loss only with respect to the first subtoken.

Training Setup As described in the main paper, all experiments are run over 5 random seeds, which are 12, 79, 237, 549, and 886.

All models were fine-tuned on a single GPU Titan X with 12 GB memory. Due to the memory restrictions of the GPUs and the dimensions of the language models, the batch size was adapted as needed (see Table A.2).

Additional Results

In this section we present addition plots that strengthen the results shown in the main paper. First, we present the results on the pre-trained language model checkpoints without any fine-tuning. Second, we present additional results of the fine-tuned language models, which are described in Chapter 5.

No fine-tuning Tables A.3 and A.4 show the prediction accuracy of the pre-trained models. Moreover, Figure A.14 shows the results of individual gaze features for all pre-trained models (without fine-tuning) on the Dundee (EN) and RSC (RU) corpora. Figure A.15 presents the differences in prediction accuracy for the pre-trained XLM-100 model predictions relative to the mean baseline for each eye tracking feature. The pre-trained models clearly cannot outperform the mean baseline for any language or dataset.

Individual feature results Figure A.17 shows the prediction accuracy for the individual eye tracking features for all datasets.

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

Table A.3: Prediction accuracy of the pre-trained language models aggregated over all eye tracking features for the English corpora, including the concatenated dataset. Standard deviation is reported in parentheses. Best results per column are marked in bold.

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

Table A.4: Prediction accuracy of the pre-trained language models aggregated over all eye tracking features for the Dutch, German and Russian corpora, and for all four languages combined in a single dataset. Standard deviation is reported in parentheses. Best results per column are marked in bold.

Predictions on parts of speech Figure A.16 shows an additional analysis where we explore which parts-of-speech can be predicted more accurately.

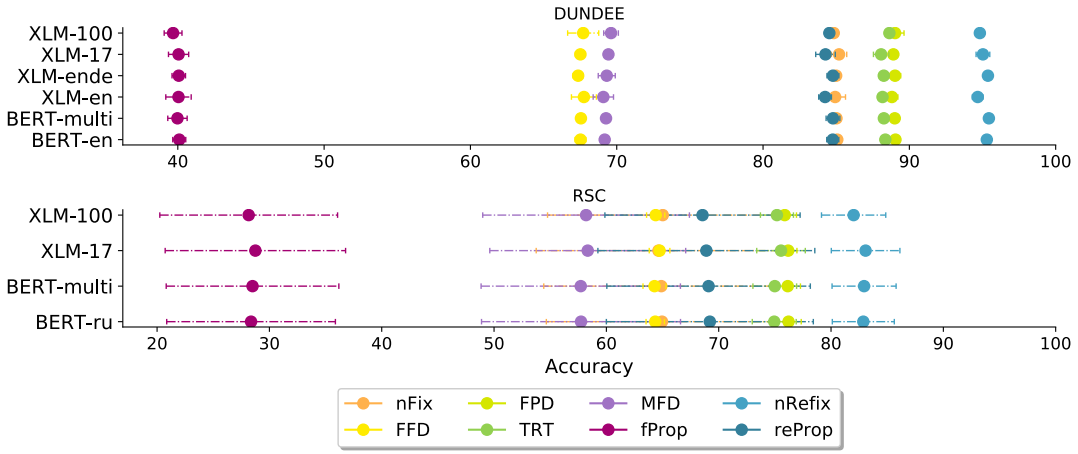


Figure A.14: Results of individual gaze features for all pre-trained models (without fine-tuning) on the Dundee (EN) and RSC (RU) corpora.

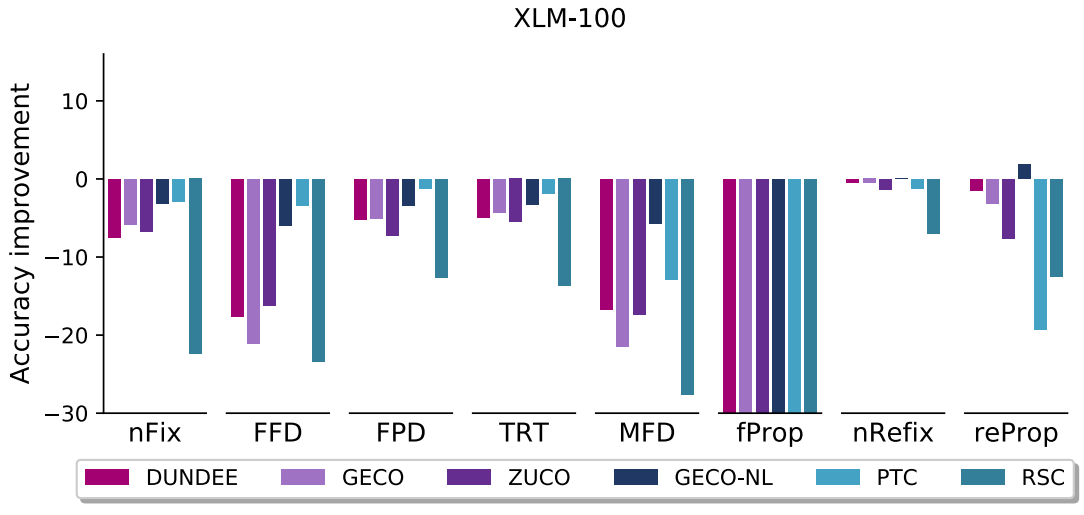


Figure A.15: Differences in prediction accuracy for the pre-trained XLM-100 model predictions relative to the mean baseline for each eye tracking feature.

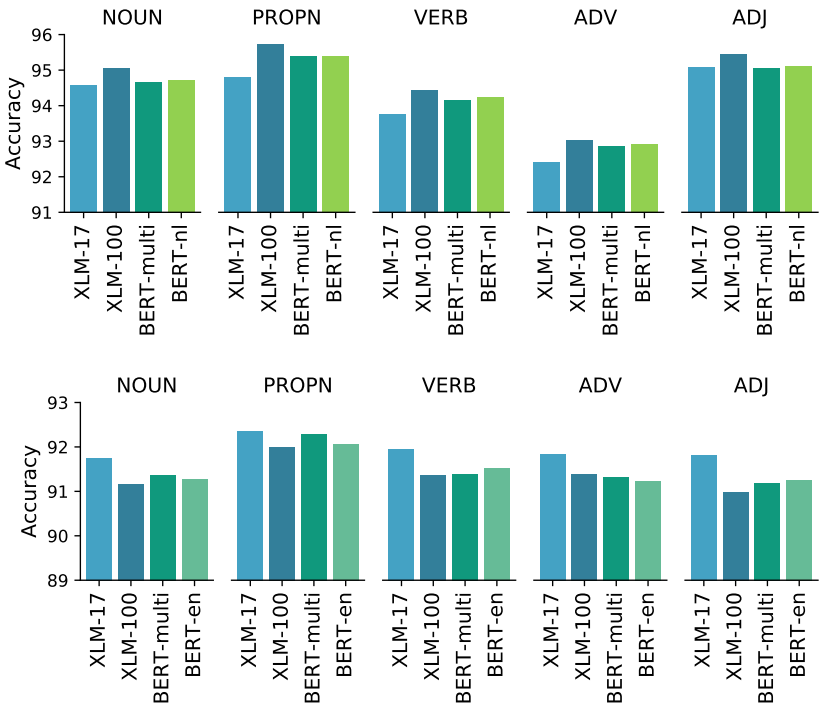


Figure A.16: Accuracy of the language models predicting the mean fixation duration (MFD) across various parts of speech for Dutch (GECO) and English (ALL-EN).

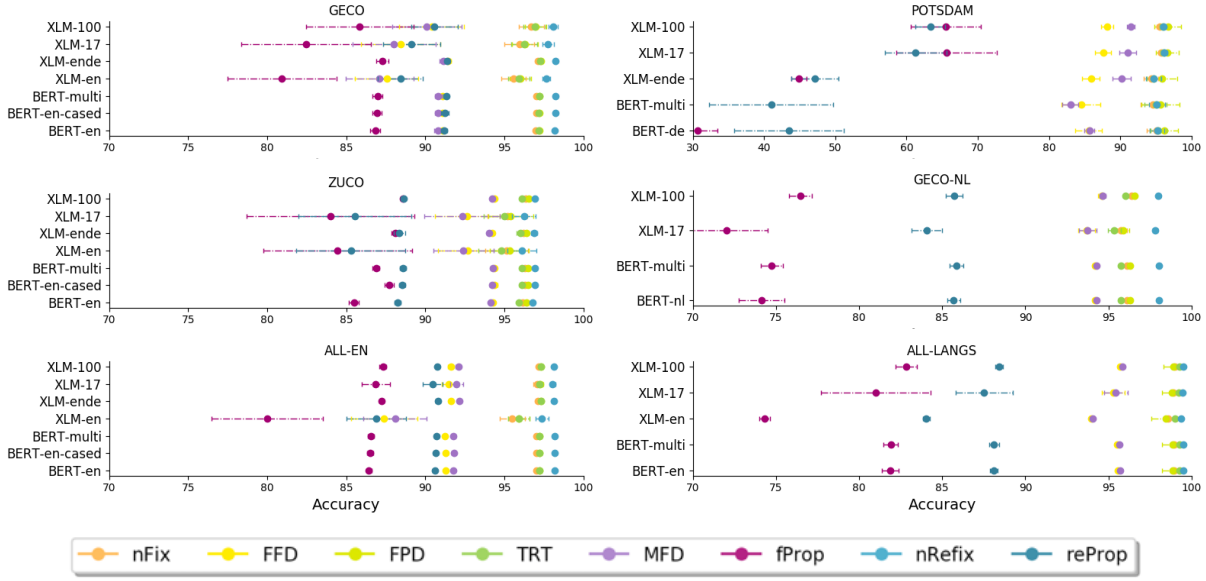


Figure A.17: Results of individual gaze features for all models on all datasets.

A.4 Correlation of CogniVal Results Between Datasets

The following plots show example correlations between the prediction results within one modality, but across datasets. These plots highlight correlations between different stimuli and different recording procedures.

Correlations between eye tracking datasets

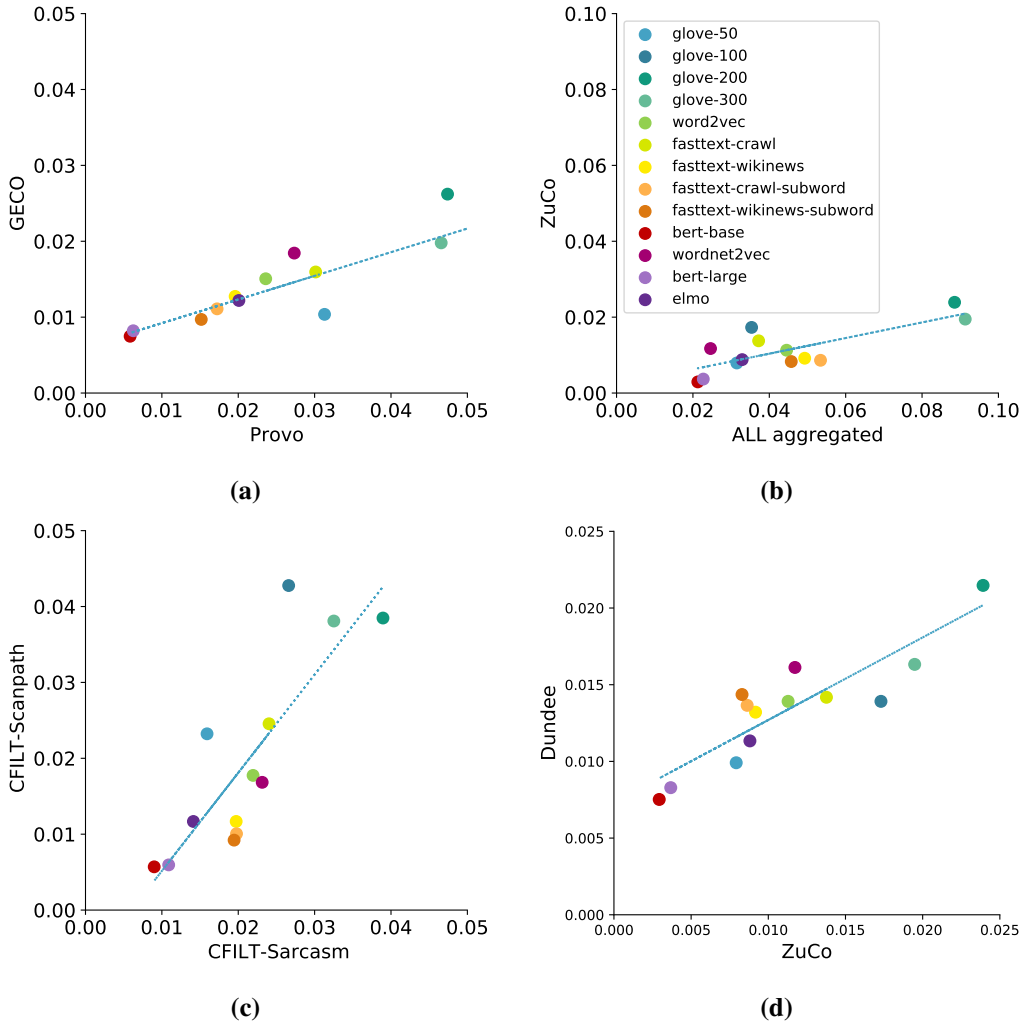


Figure A.18: Correlation plots between the prediction results of eye tracking datasets.

Figure A.18 shows the correlations of results between various eye tracking datasets. For instance, we compare the performance of the embeddings predicting eye tracking features from the Dundee corpus and from the ZuCo corpus, which include texts of different genres.

Correlations between fMRI datasets

Figure A.19 shows the correlations of the results between various fMRI datasets. For example, a clear correlation can be observed between the Alice dataset, a reading study of natural continuous reading stimuli, and the Harry Potter dataset, a controlled word-by-word reading study.

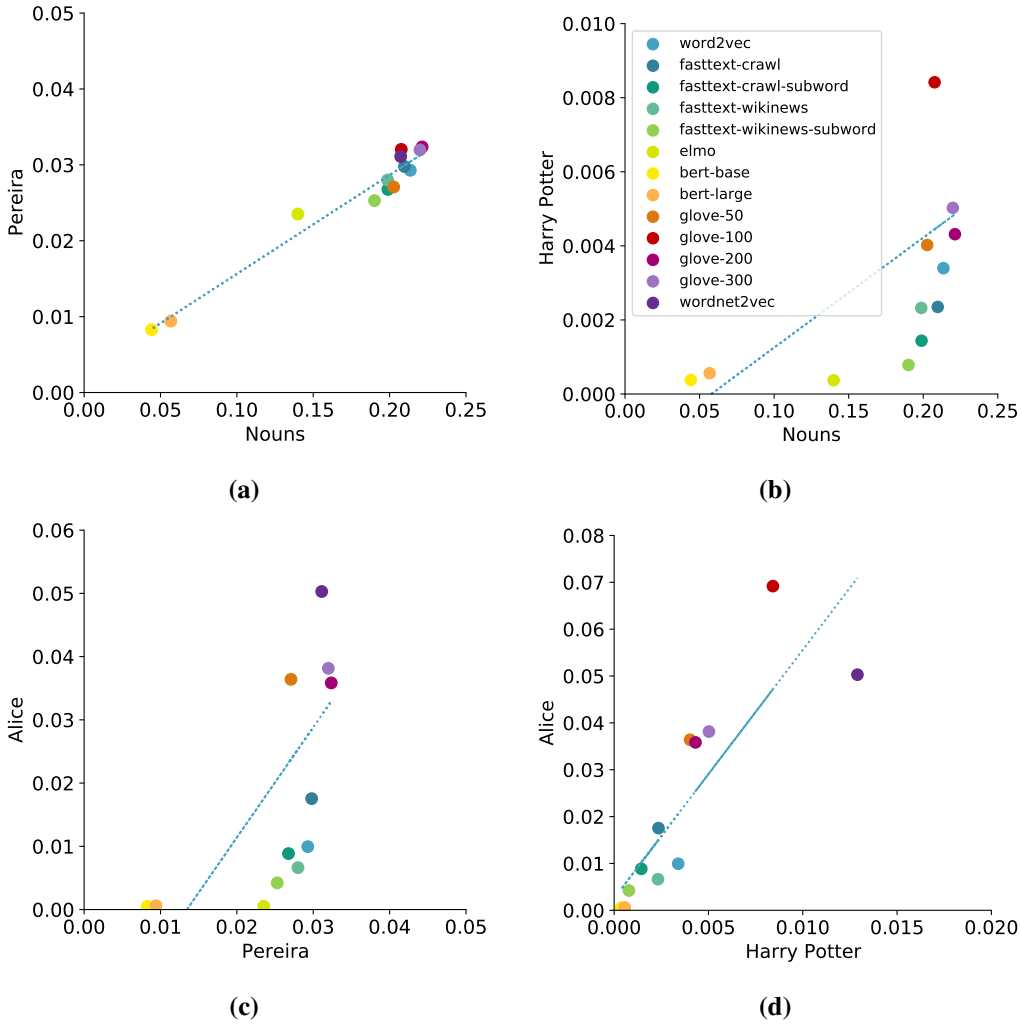


Figure A.19: Correlation plots between the prediction results of fMRI datasets.

Bibliography

Samira Abnar, Rasyan Ahmed, Max Mijnheer, and Willem Zuidema. 2018. Experiential, distributional and dependency-based word embeddings have complementary roles in decoding brain activity. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 57–66.

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

Héctor Martínez Alonso and Barbara Plank. 2017. When is multitask learning effective? semantic sequence prediction under varying data conditions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 44–53.

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

Simran Arora, Avner May, Jian Zhang, and Christopher Ré. 2020. Contextual embeddings: When are they worth it? *arXiv preprint arXiv:2005.09117*.

Ekaterina Artemova, Amir Bakarov, Aleksey Artemov, Evgeny Burnaev, and Maxim Sharaev. 2020. Data-driven models and computational tools for neurolinguistics: a language technology perspective. *Journal of Cognitive Science*, 21(1):15–52.

Nicholas A Badcock, Petroula Mousikou, Yatin Mahajan, Peter De Lissa, Johnson Thie, and Genevieve McArthur. 2013. Validation of the Emotiv EPOC® EEG gaming system for measuring research quality auditory ERPs. *PeerJ*, 1:e38.

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

Amir Bakarov. 2018a. Can eye movement data be used as ground truth for word embeddings evaluation? In *Proceedings of the 11th International Conference on Language Resources and Evaluation*.

Amir Bakarov. 2018b. A survey of word embeddings evaluation methods. *arXiv preprint arXiv:1801.09536*.

Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. 2017. Assessing state-of-the-art sentiment models on state-of-the-art sentiment datasets. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 2–12.

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

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

Maria Barrett, Ana Valeria González-Garduño, Lea Frermann, and Anders Søgaard. 2018b. Unsupervised induction of linguistic categories with records of reading, speaking, and writing. 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 2028–2038.

Maria Barrett and Anders Søgaard. 2015a. Reading behavior predicts syntactic categories. In *Proceedings of the 19th Conference on Computational Natural Language Learning*, pages 345–349.

Maria Barrett and Anders Søgaard. 2015b. Using reading behavior to predict grammatical functions. In *Proceedings of the 6th Workshop on Cognitive Aspects of Computational Language Learning*, pages 1–5.

Marcel CM Bastiaansen, Jos JA Van Berkum, and Peter Hagoort. 2002. Event-related theta power increases in the human EEG during online sentence processing. *Neuroscience Letters*, 323(1):13–16.

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

Lisa Beinborn, Samira Abnar, and Rochelle Choenni. 2019. Robust evaluation of language-brain encoding experiments. *International Journal of Computational Linguistics and Applications*.

Emily M Bender. 2018. How to make ends meet: Why general purpose NLU needs linguistics. In *Talk presented at the Workshop on Relevance of Linguistic Structure in Neural Architectures for NLP (RELNLP) at ACL*.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Joachim Bingel and Anders Søgaard. 2017. Identifying beneficial task relations for multi-task learning in deep neural networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 164–169.

Philippe Blache, Stéphane Rauzy, Deirdre Bolger, Chotiga Pattamadilok, and Sophie Dufour. 2018. A dataset for studying idiom processing with EEG. In *Linguistic and Neuro-Cognitive Resources (LiNCR), LREC 2018 Workshop*, pages 18–22.

Carlo Bonferroni. 1936. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62.

David H Brainard. 1997. The psychophysics toolbox. *Spatial Vision*, 10(4):433–436.

Jonathan R Brennan, Edward P Stabler, Sarah E Van Wagenen, Wen-Ming Luh, and John T Hale. 2016. Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and Language*, 157:81–94.

Michael P Broderick, Andrew J Anderson, Giovanni M Di Liberto, Michael J Crosse, and Edmund C Lalor. 2018. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology*, 28(5):803–809.

Andreas Bruns. 2004. Fourier-, Hilbert-and wavelet-based signal analysis: Are they really different approaches? *Journal of Neuroscience Methods*, 137(2):321–332.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

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

Rich Caruana. 1997. Multitask learning. *Machine Learning*, 28(1):41–75.

Marco Catani, Derek K Jones, and Dominic H Ffytche. 2005. Perisylvian language networks of the human brain. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 57(1):8–16.

Branden Chan, Timo Möller, Malte Pietsch, Tanay Soni, and Chin Man Yeung. German BERT.

Yun-Nung Chen, Kai-Min Chang, and Jack Mostow. 2012. Towards using EEG to improve ASR accuracy. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 382–385. Association for Computational Linguistics.

Joe Cheri, Abhijit Mishra, and Pushpak Bhattacharyya. 2016. Leveraging annotators' gaze behaviour for coreference resolution. In *Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning*, pages 22–26.

Billy Chiu, Anna Korhonen, and Sampo Pyysalo. 2016. Intrinsic evaluation of word vectors fails to predict extrinsic performance. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 1–6.

Jason PC Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Agatha Christie. 1920. *The Mysterious Affair at Styles*. Retrieved from Project Gutenberg, *www.gutenberg.org*.

Charles Clifton, Adrian Staub, and Keith Rayner. 2007. Eye movements in reading words and sentences. In *Eye Movements*, pages 341–371. Elsevier.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Uschi Cop, Nicolas Dirix, Denis Drieghe, and Wouter Duyck. 2017. Presenting GECO: An eye-tracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 49(2):602–615.

Alexander Craik, Yongtian He, and Jose L Contreras-Vidal. 2019. Deep learning for electroencephalogram (EEG) classification tasks: a review. *Journal of Neural Engineering*, 16(3):031001.

Aron Culotta, Andrew McCallum, and Jonathan Betz. 2006. Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 296–303.

Michael Dambacher and Reinhold Kliegl. 2007. Synchronizing timelines: Relations between fixation durations and N400 amplitudes during sentence reading. *Brain Research*, 1155:147–162.

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

Federica Degno and Simon P Liversedge. 2020. Eye movements and fixation-related potentials in reading: A review. *Vision*, 4(1):11.

Morteza Dehghani, Reihane Boghrati, Kingson Man, Joe Hoover, Sarah I Gimbel, Ashish Vaswani, Jason D Zevin, Mary Helen Immordino-Yang, Andrew S Gordon, Antonio Damasio,

et al. 2017. Decoding the neural representation of story meanings across languages. *Human Brain Mapping*, 38(12):6096–6106.

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

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

Olaf Dimigen, Werner Sommer, Annette Hohlfeld, Arthur M Jacobs, and Reinhold Kliegl. 2011. Coregistration of eye movements and EEG in natural reading: analyses and review. *Journal of Experimental Psychology: General*, 140(4):552.

Tien Huu Do, Duc Minh Nguyen, Evaggelia Tsiligianni, Bruno Cornelis, and Nikos Deligiannis. 2017. Multiview deep learning for predicting Twitter users' location. *arXiv preprint arXiv:1712.08091*.

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

Rotem Dror, Gili Baumer, Marina Bogomolov, and Roi Reichart. 2017. Replicability analysis for natural language processing: Testing significance with multiple datasets. *Transactions of the Association for Computational Linguistics*, 5:471–486.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392.

Paola E Dussias. 2010. Uses of eye-tracking data in second language sentence processing research. *Annual Review of Applied Linguistics*, 30:149–166.

Ariel Ephrat, Tavi Halperin, and Shmuel Peleg. 2017. Improved speech reconstruction from silent video. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 455–462.

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

Allyson Ettinger, Naomi Feldman, Philip Resnik, and Colin Phillips. 2016. Modeling N400 amplitude using vector space models of word representation. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*.

Manuel JA Eugster, Tuukka Ruotsalo, Michiel M Spapé, Ilkka Kosunen, Oswald Barral, Niklas Ravaja, Giulio Jacucci, and Samuel Kaski. 2014. Predicting term-relevance from brain signals. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 425–434. ACM.

Sarah Fabi and Hartmut Leuthold. 2018. Racial bias in empathy: Do we process dark-and fair-colored hands in pain differently? An EEG study. *Neuropsychologia*, 114:143–157.

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

Chris Foster, Dhanush Dharmaretnam, Haoyan Xu, Alona Fyshe, and George Tzanetakis. 2018. Decoding music in the human brain using eeg data. In *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE.

Stefan L Frank. 2017. Word embedding distance does not predict word reading time. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society*.

Stefan L Frank, Irene Fernandez Monsalve, Robin L Thompson, and Gabriella Vigliocco. 2013. Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods*, 45(4):1182–1190.

Stefan L Frank, Leun J Otten, Giulia Galli, and Gabriella Vigliocco. 2015. The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, 140:1–11.

Stefan L Frank and Roel M Willems. 2017. Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. *Language, Cognition and Neuroscience*, 32(9):1192–1203.

Richard Futrell, Edward Gibson, Harry J Tily, Idan Blank, Anastasia Vishnevetsky, Steven Piantadosi, and Evelina Fedorenko. 2018. The Natural Stories Corpus. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*.

Alona Fyshe, Partha P Talukdar, Brian Murphy, and Tom M Mitchell. 2014. Interpretable semantic vectors from a joint model of brain-and text-based meaning. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 489–499.

Jon Gauthier and Anna Ivanova. 2018. Does the brain represent words? An evaluation of brain decoding studies of language understanding. *arXiv preprint arXiv:1806.00591*.

Sahar Ghannay, Benoit Favre, Yannick Esteve, and Nathalie Camelin. 2016. Word embedding evaluation and combination. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, pages 300–305.

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

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

Jose Gómez-Poveda and Elena Gaudioso. 2016. Evaluation of temporal stability of eye tracking algorithms using webcams. *Expert Systems with Applications*, 64:69–83.

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

Ana Valeria González-Garduño and Anders Søgaard. 2018. Learning to predict readability using eye-movement data from natives and learners. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*.

Roland H Grabner, Clemens Brunner, Robert Leeb, Christa Neuper, and Gert Pfurtscheller. 2007. Event-related eeg theta and alpha band oscillatory responses during language translation. *Brain Research Bulletin*, 72(1):57–65.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5-6):602–610.

Thomas L Griffiths, Frederick Callaway, Michael B Chang, Erin Grant, Paul M Krueger, and Falk Lieder. 2019. Doing more with less: meta-reasoning and meta-learning in humans and machines. *Current Opinion in Behavioral Sciences*, 29:24–30.

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

Lea A Hald, Marcel CM Bastiaansen, and Peter Hagoort. 2006. Eeg theta and gamma responses to semantic violations in online sentence processing. *Brain and Language*, 96(1):90–105.

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

John Hale, Chris Dyer, Adhiguna Kuncoro, and Jonathan R Brennan. 2018. Finding syntax in human encephalography with beam search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2727–2736.

Liberty S Hamilton and Alexander G Huth. 2018. The revolution will not be controlled: Natural stimuli in speech neuroscience. *Language, Cognition and Neuroscience*, pages 1–10.

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

Olaf Hauk and Friedemann Pulvermüller. 2004. Effects of word length and frequency on the human event-related potential. *Clinical Neurophysiology*, 115(5):1090–1103.

John M Henderson, Wonil Choi, Matthew W Lowder, and Fernanda Ferreira. 2016. Language structure in the brain: A fixation-related fMRI study of syntactic surprisal in reading. *NeuroImage*, 132:293–300.

John M Henderson, Wonil Choi, Steven G Luke, and Rutvik H Desai. 2015. Neural correlates of fixation duration in natural reading: Evidence from fixation-related fMRI. *NeuroImage*, 119:390–397.

John M Henderson, Steven G Luke, Joseph Schmidt, and John E Richards. 2013. Co-registration of eye movements and event-related potentials in connected-text paragraph reading. *Frontiers in Systems Neuroscience*, 7:28.

Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3):61–83.

Agneta Herlitz and Johanna Lovén. 2013. Sex differences and the own-gender bias in face recognition: a meta-analytic review. *Visual Cognition*, 21(9-10):1306–1336.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Nora Hollenstein, Maria Barrett, and Lisa Beinborn. 2020a. Towards best practices for leveraging human language processing signals for natural language processing. In *Proceedings of the Second Workshop on Linguistic and Neurocognitive Resources*, pages 15–27.

Nora Hollenstein, Maria Barrett, Marius Troendle, Francesco Bigiolli, Nicolas Langer, and Ce Zhang. 2019a. Advancing NLP with cognitive language processing signals. *arXiv preprint arXiv:1904.02682*.

Nora Hollenstein, Adrian van der Lek, and Ce Zhang. 2020b. Cognival in action: An interface for customizable cognitive word embedding evaluation. In *Proceedings of the 28th International Conference on Computational Linguistics*.

Nora Hollenstein, Cedric Renggli, Maria Barrett, Marius Troendle, Nicolas Langer, and Ce Zhang. 2020c. A large-scale study of decoding EEG brain activity for multi-modal natural language processing. *Under Review*.

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

Nora Hollenstein, Antonio de la Torre, Nicolas Langer, and Ce Zhang. 2019b. CogniVal: A framework for cognitive word embedding evaluation. In *Proceedings of the 23nd Conference on Computational Natural Language Learning*.

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

Nora Hollenstein and Ce Zhang. 2019. Entity recognition at first sight: Improving NER with eye movement information. In *Proceedings of the 2018 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers).

Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Alexander G Huth, Wendy A de Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458.

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

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556.

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

Aditya Joshi, Abhijit Mishra, Nivvedan Senthamilselvan, and Pushpak Bhattacharyya. 2014. Measuring sentiment annotation complexity of text. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 36–41.

Salil Joshi, Diptesh Kanojia, and Pushpak Bhattacharyya. 2013. More than meets the eye: Study of human cognition in sense annotation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 733–738.

Barbara J Juhasz and Keith Rayner. 2003. Investigating the effects of a set of intercorrelated variables on eye fixation durations in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(6):1312.

Marcel A Just and Patricia A Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4):329.

Katerina D Kandylaki and Ina Bornkessel-Schlesewsky. 2019. From story comprehension to the neurobiology of language. *Language, Cognition and Neuroscience*, 34(4):405–410.

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

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

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

Douwe Kiela, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2018. Efficient large-scale multi-modal classification. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Adam Kilgarriff. 1995. BNC database and word frequency lists. *Retrieved Dec. 2017*.

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

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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

Sigrid Klerke and Barbara Plank. 2019. At a glance: The impact of gaze aggregation views on syntactic tagging. In *Proceedings of the Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)*, pages 51–61.

Wolfgang Klimesch. 2012. Alpha-band oscillations, attention, and controlled access to stored information. *Trends in Cognitive Sciences*, 16(12):606–617.

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

Marta Kutas and Kara D Federmeier. 2000. Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, 4(12):463–470.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. 2015. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270.

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

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

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Kristin Lemhöfer and Mirjam Broersma. 2012. Introducing LexTALE: A quick and valid lexical test for advanced learners of english. *Behavior Research Methods*, 44(2):325–343.

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

Mu Li and Bao-Liang Lu. 2009. Emotion classification based on gamma-band EEG. In *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, pages 1223–1226. IEEE.

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

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open Sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253.

Shouyu Ling, Andy CH Lee, Blair C Armstrong, and Adrian Nestor. 2019. How are visual words represented? Insights from EEG-based visual word decoding, feature derivation and image reconstruction. *Human Brain Mapping*, 40(17):5056–5068.

Linguistic Data Consortium. 2005. ACE (Automatic Content Extraction) English annotation guidelines for entities. *Version*, 5(6):2005–08.

Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217.

Yunfei Long, Qin Lu, Rong Xiang, Minglei Li, and Chu-Ren Huang. 2017. A cognition based attention model for sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 462–471.

Alessandro Lopopolo, Stefan L Frank, Antal Van den Bosch, Annabel Nijhof, and Roel M Willems. 2018. The Narrative Brain Dataset (NBD), an fMRI dataset for the study of natural language processing in the brain. In *LREC 2018 Workshop on Linguistic and Neuro-Cognitive Resources (LiNCR)*. LREC.

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

Steven G Luke and Kiel Christianson. 2017. The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, pages 1–8.

Björn Lundquist and Øystein A Vangsnes. 2018. Language separation in bidialectal speakers: Evidence from eye tracking. *Frontiers in Psychology*, 9:1394.

Phan Luu and Thomas Ferree. 2005. Determination of the HydroCel Geodesic Sensor Nets' average electrode positions and their 10–10 international equivalents. *Inc, Technical Note*, pages 1–11.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 1:1064–1074.

Silvia Makowski, Lena A Jäger, Ahmed Abdelwahab, Niels Landwehr, and Tobias Scheffer. 2018. A discriminative model for identifying readers and assessing text comprehension from eye movements. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 209–225. Springer.

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

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

Viorica Marian, Michael Spivey, and Joy Hirsch. 2003. Shared and separate systems in bilingual language processing: Converging evidence from eyetracking and brain imaging. *Brain and Language*, 86(1):70–82.

Sandeep Mathias, Diptesh Kanojia, Abhijit Mishra, and Pushpak Bhattacharyya. 2020. A survey on using gaze behaviour for natural language processing. *Proceedings of IJCAI*.

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

James L McClelland, Felix Hill, Maja Rudolph, Jason Baldridge, and Hinrich Schütze. 2020. Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. *Proceedings of the National Academy of Sciences*.

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

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

Francis M Miezin, L Maccotta, JM Ollinger, SE Petersen, and RL Buckner. 2000. Characterizing the hemodynamic response: effects of presentation rate, sampling procedure, and the possibility of ordering brain activity based on relative timing. *NeuroImage*, 11(6):735–759.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

George A Miller and Christiane Fellbaum. 1992. WordNet and the organization of lexical memory. In *Intelligent tutoring systems for foreign language learning*, pages 89–102. Springer.

Gosse Minnema and Aurélie Herbelot. 2019. From brain space to distributional space: the perilous journeys of fMRI decoding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 155–161.

Abhijit Mishra and Pushpak Bhattacharyya. 2018. *Cognitively Inspired Natural Language Processing: An Investigation Based on Eye-tracking*. Springer.

Abhijit Mishra, Diptesh Kanojia, and Pushpak Bhattacharyya. 2016a. Predicting readers' sarcasm understandability by modeling gaze behavior. In *AAAI Conference on Artificial Intelligence*, pages 3747–3753.

Abhijit Mishra, Diptesh Kanojia, Seema Nagar, Kuntal Dey, and Pushpak Bhattacharyya. 2016b. Harnessing cognitive features for sarcasm detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1095–1104.

Abhijit Mishra, Diptesh Kanojia, Seema Nagar, Kuntal Dey, and Pushpak Bhattacharyya. 2017a. Leveraging cognitive features for sentiment analysis. *Proceedings of The 20th Conference on Computational Natural Language Learning*, pages 156–166.

Abhijit Mishra, Diptesh Kanojia, Seema Nagar, Kuntal Dey, and Pushpak Bhattacharyya. 2017b. Scanpath complexity: Modeling reading effort using gaze information. In *Thirty-First AAAI Conference on Artificial Intelligence*, pages 4429–4436.

Tom M Mitchell, Svetlana V Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L Malave, Robert A Mason, and Marcel Adam Just. 2008. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195.

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

Jack Mostow, Kai-min Chang, and Jessica Nelson. 2011. Toward exploiting EEG input in a reading tutor. In *Artificial Intelligence in Education*, pages 230–237. Springer.

Christoph Mulert. 2013. Simultaneous EEG and fMRI: Towards the characterization of structure and dynamics of brain networks. *Dialogues in Clinical Neuroscience*, 15(3):381.

Matthias M Müller, Andreas Keil, Thomas Gruber, and Thomas Elbert. 1999. Processing of affective pictures modulates right-hemispheric gamma band eeg activity. *Clinical Neurophysiology*, 110(11):1913–1920.

Brian Murphy and Massimo Poesio. 2010. Detecting semantic category in simultaneous EEG/MEG recordings. In *Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics*, pages 36–44. Association for Computational Linguistics.

Brian Murphy, Leila Wehbe, and Alona Fyshe. 2018. Decoding language from the brain. *Language, Cognition, and Computational Models*, page 53.

Lukas Muttenthaler, Nora Hollenstein, and Maria Barrett. 2020. Human brain activity for machine attention. *arXiv preprint arXiv:2006.05113*.

Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. 2011. Encoding and decoding in fMRI. *NeuroImage*, 56(2):400–410.

Neha Nayak, Gabor Angeli, and Christopher D Manning. 2016. Evaluating word embeddings using a representative suite of practical tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 19–23.

Malvina Nissim, Rik van Noord, and Rob van der Goot. 2020. Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, 46(2):487–497.

Ewan Nurse, Benjamin S Mashford, Antonio Jimeno Yepes, Isabell Kiral-Kornek, Stefan Harrer, and Dean R Freestone. 2016. Decoding EEG and LFP signals using deep learning: Heading TrueNorth. In *Proceedings of the ACM International Conference on Computing Frontiers*, pages 259–266.

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

George Panagopoulos. 2017. Multi-task learning for commercial brain computer interfaces. In *2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 86–93. IEEE.

Alexandra Papoutsaki, Patsorn Sangkloy, James Laskey, Nediyana Daskalova, Jeff Huang, and James Hays. 2016. WebGazer: Scalable webcam eye tracking using user interactions. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence-IJCAI 2016*.

Lucas C Parra, Clay D Spence, Adam D Gerson, and Paul Sajda. 2005. Recipes for the linear analysis of EEG. *Neuroimage*, 28(2):326–341.

Andreas Pedroni, Amirreza Bahreini, and Nicolas Langer. 2019. Automagic: Standardized preprocessing of big EEG data. *NeuroImage*.

Denis G Pelli. 1997. The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, 10(4):437–442.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

Daniela Perani, Stanislas Dehaene, Franco Grassi, Laurent Cohen, Stefano F Cappa, Emmanuel Dupoux, Ferruccio Fazio, and Jacques Mehler. 1996. Brain processing of native and foreign

languages. *NeuroReport-International Journal for Rapid Communications of Research in Neuroscience*, 7(15):2439–2444.

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. 2018. Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, 9(1):963.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237.

Christian Pfeiffer, Nora Hollenstein, Ce Zhang, and Nicolas Langer. 2020. Neural dynamics of sentiment processing during naturalistic sentence reading. *NeuroImage*, page 116934.

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

Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in NLP. *KONVENS*.

Cathy J Price. 2012. A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. *NeuroImage*, 62(2):816–847.

Yanina Prystauka and Ashley Glen Lewis. 2019. The power of neural oscillations to inform sentence comprehension: A linguistic perspective. *Language and Linguistics Compass*, 13(9):e12347.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Dhanesh Ramachandram and Graham W Taylor. 2017. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6):96–108.

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

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

Keith Rayner and Susan A Duffy. 1986. Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, 14(3):191–201.

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

Joao António Rodrigues, Ruben Branco, João Silva, Chakaveh Saedi, and António Branco. 2018. Predicting brain activation with WordNet embeddings. In *Proceedings of the Eight Workshop on Cognitive Aspects of Computational Language Learning and Processing*, pages 1–5.

Anna Rogers, Shashwath Hosur Ananthakrishna, and Anna Rumshisky. 2018. What's in your embedding, and how it predicts task performance. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2690–2703.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Omid Rohanian, Shiva Taslimipoor, Victoria Yaneva, and Le An Ha. 2017. Using gaze data to predict multiword expressions. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 601–609.

Jonathan Rotsztejn, Nora Hollenstein, and Ce Zhang. 2018. ETH-DS3Lab at SemEval-2018 Task 7: Effectively combining recurrent and convolutional neural networks for relation classification and extraction. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 689–696.

Sebastian Ruder, Joachim Bingel, Isabelle Augenstein, and Anders Søgaard. 2017. Learning what to share between loosely related tasks. *arXiv preprint arXiv:1705.08142*.

G Rupert Jr et al. 2012. *Simultaneous statistical inference*. Springer Science & Business Media.

Chakaveh Saedi, António Branco, João António Rodrigues, and João Silva. 2018. WordNet embeddings. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 122–131.

Javier San Agustin, Henrik Skovsgaard, John Paulin Hansen, and Dan Witzner Hansen. 2009. Low-cost gaze interaction: Ready to deliver the promises. In *CHI'09 Extended Abstracts on Human Factors in Computing Systems*, pages 4453–4458.

Martin Schrimpf, Idan A Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy G Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2020. Artificial neural networks accurately predict language processing in the brain. *BioRxiv*.

Dan Schwartz and Tom Mitchell. 2019. Understanding language-elicited EEG data by predicting it from a fine-tuned language model. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 43–57.

Dan Schwartz, Mariya Toneva, and Leila Wehbe. 2019. Inducing brain-relevant bias in natural language processing models. In *Advances in Neural Information Processing Systems*, pages 14100–14110.

Sara C Sereno and Keith Rayner. 2003. Measuring word recognition in reading: Eye movements and event-related potentials. *Trends in Cognitive Sciences*, 7(11):489–493.

Weston Sewell and Oleg Komogortsev. 2010. Real-time eye gaze tracking with an unmodified commodity webcam employing a neural network. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, pages 3739–3744.

Cory Shain, Idan Asher Blank, Marten van Schijndel, William Schuler, and Evelina Fedorenko. 2019. fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*.

David L Share. 2008. On the anglocentricities of current reading research and practice: The perils of overreliance on an "outlier" orthography. *Psychological Bulletin*, 134(4):584.

Abhinav Deep Singh, Poojan Mehta, Samar Husain, and Rajkumar Rajakrishnan. 2016. Quantifying sentence complexity based on eye-tracking measures. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (cl4lc)*, pages 202–212.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.

Anders Søgaard. 2016. Evaluating word embeddings with fMRI and eye-tracking. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 116–121.

Ekta Sood, Simon Tannert, Diego Frassinelli, Andreas Bulling, and Ngoc Thang Vu. 2020a. Interpreting attention models with human visual attention in machine reading comprehension. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 12–25.

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

Brigitte Stemmer and John F Connolly. 2012. The EEG/ERP technologies in linguistic research. *Methodological and Analytic Frontiers in Lexical Research*, 47:337.

Kurt Stocker and Matthias Hartmann. 2019. "Next Wednesday's meeting has been moved forward two days": The time-perspective question is ambiguous in Swiss German, but not in Standard German. *Swiss Journal of Psychology*, 78(1-2):61.

Michalina Strzyz, David Vilares, and Carlos Gómez-Rodríguez. 2019. Towards making a dependency parser see. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1500–1506.

Kirill Stytsenko, Evaldas Jablonskis, and Cosima Prahm. 2011. Evaluation of consumer EEG device Emotiv EPOC. In *MEi: CogSci Conference 2011, Ljubljana*.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune BERT for text classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer.

Simon Suster, Stéphan Tulkens, and Walter Daelemans. 2017. A short review of ethical challenges in clinical natural language processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 80–87.

Ece Takmaz, Sandro Pezzelle, Lisa Beinborn, and Raquel Fernández. 2020. Generating image descriptions via sequential cross-modal alignment guided by human gaze. In *Proceedings*

of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4664–4677.

Erik F Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the 7th Conference on Natural Language Learning*, volume 4, pages 142–147.

Takenobu Tokunaga, Hitoshi Nishikawa, and Tomoya Iwakura. 2017. An eye-tracking study of named entity annotation. *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 758–764.

Katrin Tomanek, Udo Hahn, Steffen Lohmann, and Jürgen Ziegler. 2010. A cognitive cost model of annotations based on eye-tracking data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1158–1167.

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

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

Lorenzo Vignali, Nicole A Himmelstoss, Stefan Hawelka, Fabio Richlan, and Florian Hutzler. 2016. Oscillatory brain dynamics during sentence reading: a fixation-related spectral perturbation analysis. *Frontiers in Human Neuroscience*, 10:191.

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

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

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

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

Sabine Weiss and Horst M Mueller. 2003. The contribution of EEG coherence to the investigation of language. *Brain and Language*, 85(2):325–343.

Sabine Weiss and Horst M Mueller. 2012. "Too many betas do not spoil the broth": the role of beta brain oscillations in language processing. *Frontiers in Psychology*, 3:201.

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

Chad C Williams, Mitchel Kappen, Cameron D Hassall, Bruce Wright, and Olave E Krigolson. 2019. Thinking theta and alpha: Mechanisms of intuitive and analytical reasoning. *NeuroImage*, 189:574–580.

Rihana Williams and Robin Morris. 2004. Eye movements, word familiarity, and vocabulary acquisition. *European Journal of Cognitive Psychology*, 16(1-2):312–339.

Irene Winkler, Stephanie Brandl, Franziska Horn, Eric Waldburger, Carsten Allefeld, and Michael Tangermann. 2014. Robust artifactual independent component classification for BCI practitioners. *Journal of Neural Engineering*, 11(3):035013.

Irene Winkler, Stefan Haufe, and Michael Tangermann. 2011. Automatic classification of artifactual ICA-components for artifact removal in EEG signals. *Behavioral and Brain Functions*, 7(1):30.

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

Esther Xiu Wen Wu, Bruno Laeng, and Svein Magnussen. 2012. Through the eyes of the own-race bias: Eye-tracking and pupillometry during face recognition. *Social neuroscience*, 7(2):202–216.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Pingmei Xu, Krista A Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R Kulkarni, and Jianxiong Xiao. 2015. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*.

Victoria Yaneva, Richard Evans, Ruslan Mitkov, et al. 2018. Classifying referential and non-referential it using gaze. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4896–4901.

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