Universität Zürich[UZH]

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

CogniValSent

An extension of the CogniVal framework to evaluate the cognitive plausibility of sentence-level embeddings

Author: Adrian van der Lek
Student ID: 09-915-422

Examiner: Prof. Dr. Martin Volk

Supervisor: Nora Hollenstein (DS3Lab, ETH Zurich)

Department of Computational Linguistics

Submission date: 01.07.2020

Abstract

In order to assess different approaches to obtaining word and sentence embedding vectors, and to form a basis upon which new methods can developed, formal evaluation is necessary. Since their inception, embeddings have been evaluated using *extrinsic* means, i.e. in downstream tasks, which serve as proxies to real world applications. More recently, *intrinsic* evaluation methods have been proposed, which strive to investigate inherent properties of embedding vectors, but typically only investigate very specific phenomena and can be subject to individual bias. An alternative is to leverage processes occurring in the human brain whilst reading or processing speech, in order to obtain a measure of the cognitive plausiblity of embedding approaches. Hollenstein et al. [2019] have shown that word embedding vectors can be used in a neural regression setting, where embeddings predict cognitive signals aggregated on the word-level. Such signals can be obtained through recordings of physiological monitoring methods such as eye-tracking, EEG and fMRI. Tested approaches differ significantly in how well they predict cognitive signals and rankings correlate between datasets, modalities, as well as with results of extrinsic evaluations. Furthermore, the authors find many results to be significantly different from randomly generated baselines. In this thesis, I apply the approach by Hollenstein et al. [2019] to sentence embeddings and sentence-level cognitive signals, with necessary adaptations. I evaluate eight sentence embedding approaches of varying complexity and select a subset of the cognitive datasets evaluated by Hollenstein et al. [2019], which offer sufficient data on the sentence level. Between approaches, I observe distinct rankings, which differ considerably between the modalities eye-tracking and EEG. For fMRI, results provide limited information, which I mainly attribute to data sparsity. *Skip-Thought* [Kiros et al., 2015] and *InferSent* [Conneau et al., 2017] stand out across modalities, yielding the lowest overall errors. Unexpectedly, two more recent approaches, *ELMo* [Peters et al., 2018] and particularly *BERT* [Devlin et al., 2018] do not reflect observed downstream performance. Lastly, I informally assess correlation between cognitive and previous intrinsic and extrinsic evaluation results. Results point toward a potential relationship between EEG and tasks measuring semantic relatedness and textual similarity, and to a lesser extent, between eye-tracking and linguistic probing tasks. To adress some of the observed issues, future work could assess different means of obtaining sentence representations, (particularly for BERT), as well as alternative regression models. As more datasets become available, more fine-grained analyses and more robust estimates of embedding performance will be possible. Finally, cognitive datasets in languages other than English open up the possibility for multi-lingual studies.

Zusammenfassung

Um Ansätze zur Gewinnung von Wort- und Satzeinbettungen zu bewerten und eine Grundlage zu schaffen, auf der neue Methoden entwickelt werden können, ist es erforderlich, formale Evaluationen durchzuführen. Seit ihrer Einführung wurden Einbettungen mit *extrinsischen* Mitteln evaluiert, d.h. mittels nachgelagerter Aufgaben, die stellvertretend für reale Anwendungen herangezogen werden. In den vergangenen Jahren wurden *intrinsische* Methoden vorgeschlagen, die inhärente Eigenschaften von Einbettungen untersuchen, sich jedoch auf spezifische Phänomene beschränken und einer subjektiven Verzerrung unterliegen können. Eine Alternative ist, Messungen von Prozessen zu nutzen, die im menschlichen Gehirn beim Lesen oder Verarbeiten von Sprache ablaufen, um die kognitive Plausibilität von Einbettungsverfahren zu beurteilen. Hollenstein et al. [2019] zeigen, dass Worteinbettungen im Rahmen einer neuronalen Regression kognitive Signale vorhersagen können. Dabei handelt es sich um Daten, die mittels der physiologischer Messverfahren Eye-Tracking, EEG und fMRI gewonnen werden. Die Rangordnung der Performanz korreliert zwischen den Datensätzen, Modalitäten, sowie mit extrinsischen Evaluationsresultaten. Zudem stellen die Autoren fest, dass sich viele Ergebnisse signifikant von zufällig generierten Basislinien unterscheiden. In dieser Arbeit wende ich den Ansatz von Hollenstein et al. [2019] angepasst auf Satzeinbettungen und Signale auf Satzebene an. Zu diesem Zweck evaluiere ich acht Ansätze anhand einer Teilmenge der von Hollenstein et al. [2019] ausgewerteten Datensätzen, die ausreichende Daten auf der Satzebene bieten. Sowohl zwischen Ansätzen als auch zwischen EEG- und fMRI-Daten beobachte ich erhebliche Unterschiede. Die fMRI-Resultate sind weniger aussagekräftig, was ich hauptsächlich auf Datenknappheit zurückführe. *Skip-Thought* [Kiros et al., 2015] und *InferSent* [Conneau et al., 2017] weisen gesamthaft die geringsten Fehler auf. Überraschend widerspiegeln die Resultate nicht die anwendungsorientierte Performanz zweier neueren Ansätze, *ELMo* [Peters et al., 2018] und *BERT* [Devlin et al., 2018], was die Erfordernis weiterführender Untersuchungen zeigt. Abschliessend bewerte ich informell die Korrelation zwischen kognitiven und intrinsischen und extrinsischen Evaluationsergebnissen. Die Ergebnisse deuten in erster Linie auf eine mögliche Beziehung zwischen EEG-Daten und Aufgaben zur Bestimmung semantischer Verwandtschaft und textueller Ähnlichkeit hin. Zukünftige Arbeiten können alternative Vorverarbeitungsmethoden (insbesondere für ELMo und BERT) sowie ggf. andere Regressionsmodelle überprüfen. Mit zunehmender Zahl an kognitiven Datensätzen werden Detailanalysen und robustere Performanzschätzungen möglich. Zuletzt schaffen kognitive Korpora in weiteren Sprachen eine Grundlage für multilinguale Erweiterungen.

Acknowledgement

I hereby thank Nora Hollenstein and Prof. Dr. Martin Volk for supervising and reviewing the present thesis. Further thanks go to Dr. Manfred Klenner, the Department of Computational Linguistics at the University of Zurich and the DS3Lab at ETH Zurich for arranging time and resources on their respective computational infrastructures. Finally, I thank my parents for their support during the creation of this thesis.

Contents

List of Figures

List of Tables

List of Acronyms

BoW	Bag of Word
CLI	Command-line interface
DNN	Deep neural network
EEG	Electroencephalography
ERP	Event-related potential
fMRI	functional Magnetic Resonance Imaging
MLP	Multi-layer perceptron
MSE	Mean Squared Error
NER	Named Entity Recognition
NLP	Natural Language Processing
POS	Part-Of-Speech

1 Introduction

The innovation of embeddings has greatly contributed to the state of the art of natural language processing in recent years. Modern embeddings are distributed representations of textual units, typically words or sentences. As such, their meaning is encoded in the dimensions of the corresponding vector, but also in dimensions of other vectors. Dimensions are not interpretable (contrasting with distributional representations where dimensions represent contexts of occurrence). A concept can be represented by multiple dimensions and conversely, a single dimension can encode information relating to multiple meanings [Goldberg, 2017]. More specifically, each unit is represented as a dense, low-dimensional vector encoding a *pattern of activation values*, which capture its meaning. The relations between units are determined by the similarities of their respective vectors [Goldberg, 2017]. Embeddings have displaced sparse, high-dimensional or *one-hot* representations of word frequencies and other manually engineered features, which suffer from what is commonly referred to as the curse of dimensionality in the context supervised learning, leading to data sparsity.

Vectors are typically obtained by learning context information from a large corpus of documents by means of un- or self-supervised learning[1], however, some approaches also leverage supervised data. Models are primarily evaluated by extrinsic means, i.e. in common downstream tasks such as classification or question answering. More recently, intrinsic evaluation approaches assessing linguistic aspects have been proposed (see e.g. Conneau and Kiela [2018]). However, as noted by Hollenstein et al. [2019], conscious judgements of linguistic properties are prone to subjective bias and are generally not predictive of extrinsic performance. On the sentence level, linguistic probing tasks as present in evaluation tools such as *SentEval*, test for a battery of narrowly defined and relatively simple tasks such as the detection of shifted bigrams or tense prediction. Recent work (e.g Søgaard [2016], Abnar et al. [2017], Schwartz and Mitchell [2019] and Hollenstein et al. [2019]) has sought to leverage recordings of brain activity, in order to relate embeddings directly to the full range

[1] In keeping with cited literature, I will use the term *unsupervised* when discussing individual approaches, however I note that the term *self-supervised* is technically more accurate and sees use in literature, e.g. [Perone et al., 2018].

of cognitive processes, as far as they can be captured by the respective approaches. Of particular interest are recordings of brain activity when subjects process text in a natural settings. This is achieved by self-paced reading, or listening to sentences or longer contiguous text segments such as books or chapters thereof. This allows the most bias-free access to lexical representations in the human brain (e.g Søgaard [2016]). As studies encompass multiple subjects, aggregate representations can be obtained that emphasize inter-subject commonalities likely to correspond to fundamental cognitive processes independent of subjective differences [2].

This thesis continues the work of [Hollenstein et al., 2019], in an effort to evaluate if and how well sentence embeddings can predict various types of sentence-level cognitive data.

1.1 Research questions

The following research questions form the basis of this thesis:

1. Can cognitive signals aggregated on the sentence-level be predicted by sentence embeddings in a neural regression setting? In particular, are results significantly different from a randomly generated baseline?
2. Do performance rankings of sentence embeddings in the present evaluation correspond to rankings in intrinsic and extrinsic evaluations?
3. There are differences of varying degree between other types of evaluation. Which are best aligned with the cognitive prediction scores?
4. How do the extrinsically best-performing sentence embeddings compare between modality?

1.2 Structure of the thesis

In this first chapter, a brief overview was given on the subject of intrinsic embedding evaluation by means of cognitive signals. Chapter 2 summarizes the various modalities (types) of cognitive signals and the means of their collection in the context of human text processing. Chapter 3 gives an overview of previous work exploring the relationship between cognitive signals, language and (artificial) neural networks, as

[2] Given the very small number and non-random selection of subjects in many studies, I note that group-related biases may remain manifest in the data.

well as examples of cognitively informed NLP. The second half introduces various embedding evaluation approaches, contrasting *extrinsic* with *intrinsic* evaluation modes. It summarizes Hollenstein et al. [2019], where the authors propose an intrinsic cognitive evaluation framework which is the foundation of this thesis. Chapter 4 describes the cognitive datasets considererd for the present evaluation, briefly summarizing their curation and general properties. Chapter 5 gives an overview over the sentence embedding approaches I choose to evaluate against the cognitive dataset. Chapter 6 outlines the evaluation setup and describes quantitative and qualitative results. Chapter 7 discusses potential confounding factors and limitations. Finally, Chapter 8 offers a conclusion and outlook as to potential future developments. A description of the command-line tool developed in the context of this thesis can be found in Appendix A.

2 Cognitive processes and language

2.1 Human language data and NLP

Mishra and Bhattacharyya [2018] state that "text is a manifestation of thought and emotion that give rise to cognitive processes in the brain. When a reader reads a piece of text, she experiences emotions, stances, nuances, subtleties, inferences, suggestions, and much more. [...] Text reveals its secrets to a willing reader, and the reader responds by moving or staying the eye, producing brain waves and making face and body movements, all of which are capturable by the modern-day technology of eye-trackers, EEGs, and MEGs." (p. vi)

Hollenstein et al. [2020] note the multiple uses of human cognitive language data. It can be used to improve NLP tasks such as part-of-speech tagging (e.g. Barrett et al. [2016]), or sentence compression (e.g. Klerke et al. [2016]). On the other hand, it can be leveraged to evaluate NLP components such as word embeddings, as done by [Hollenstein et al., 2019] and continued in this work.

Following, I briefly discuss aspects of speech perception and reading and the cognitive processes, or brain activity, triggered by these processes. I continue by summarizing three major methods of recording this brain activity. Such recordings allow to curate the cognitive datasets that form the basis of cognitive embedding evaluation.

2.2 Speech perception and reading

Speech perception begins with sound waves representing speech entering the ear of a listener. The brain extracts speech from sound waves, which in turn activate word meanings. Processing steps occur at different points in time, ranging from acoustic-phonetic features (50-100ms), language-specific phonetic-phonological analysis (100-200ms) and lexical-semantic activation (200ms onwards), the latter being of interest in the scope of this work [Salmelin, 2007].

In the case of *reading*, the dual-route model [Coltheart et al., 1993] states that fa-

miliar and unfamiliar words are processed differently. Familliar ones are assumed to be processed first at the level of single letters, then as whole word, which activates its meaning and sound form. For unfamiliar, graphemes are instead mapped to a phonological representation, which does not necessarily lead to a semantic association. Across time, basic visual feature analysis occurs in the brain at 100 milliseconds (ms), and the analysis of letters is observed at 150 ms in the visual cortex. At 200-600 ms, the temporal lobe shows activation related to reading comprehension, distinguishing words from non-words [Salmelin, 2007]. Humans read at an average speed of three words per second and the reading process involves perceiving and gradual integration of incoming words in order to obtain a representation of meaning [Wehbe et al., 2014b].

2.3 Methods of monitoring cognitive processes related to language

Researching cognitive processes relating to language processing has a long-standing tradition in psycholinguistics, with first mechanical means of measuring signals appearing at the start of the 20th century. Following, I summarize three main methods for recording cognitive signals over time which are in current use: *Eye-tracking*, an indirect approach monitoring eye-movement with respect to a visual stimulus; *electroencephalography* (EEG), which records electrical brain activity relating to a stimulus through the scalp and also captures spatial aspects of the brain, and finally *functional Magnetic Resonance Imaging* (fMRI), which fully maps all three dimensions of the brain and allows to precisely pinpoint locations of activity triggered by a stimulus.

This brief overview already illustrates the increasing complexity and expressivity of these techniques. Furthermore, and somewhat trivially, eye-tracking only operates on written text in the context of language research, whilst EEG and fMRI can be used to study both text and speech perception. However, I will highlight throughout this thesis that both eye-tracking and EEG remain relevant for both intrinsic and practical reasons. Specifically, the three methods constitute complementary *modalities* in the context of cognitive embedding evaluation.

2.3.1 Eye-Tracking

Eye-tracking is the oldest of the presently discussed techniques, with approaches relying on purely mechanical means such as specialized contact-lenses appearing at the the beginning of the 20th century (e.g. Huey [1908]). It is also the most indirect of the three methods discussed here. Modern eye-tracking captures the focus of a viewer's gaze on a (visual) stimulus over time [Mishra and Bhattacharyya, 2018]. When reading, eye movements are on one hand influenced by low-level factors such as what the eye captures during each fixation and the length of a word. On the other hand, they are also determined by high-level factors such as syntactic processing, with low- and high-level factors interrelating in a complex way [Barrett et al., 2015]. In the 1980s, systems were developed that could be operated with personal computers, were capable of high-speed data processing and could be used as "as interface to facilitate interaction between humans and computer."[Mishra and Bhattacharyya, 2018][22]. Cop et al. [2017][603] note that "[w]ith modern-day eyetracking [sic] equipment, the position of the eye can be determined every millisecond with very high spatial accuracy, resulting in a very rich and detailed dataset". They further explain that "[...] when the goal is to explain how reading occurs in natural contexts, the ambition of reading models should also be to expand their generalizability beyond word-level processes, in order to cover a larger scope of potential interacting language processes." In particular, they emphasize the importance of considering the interaction between word-level and semantic or syntactic processes, which may be observed when readers process longer text segments. Cop et al. [2017] further note that eye-tracking allows observation of silent reading and minimizes the need for instruction or intervention on the researcher's part. I note in line with [Hollenstein et al., 2019] that this allows self-paced, natural reading, which is particularly beneficial for assessing sentence embeddings in the context of cognitive evaluation.

Typically, both eye and head positions need to be measured to determine the exact location of the viewer's gaze. Both invasive and non-invasive methods exist, with invasive approaches requiring the insertion of objects in the eye, but generally providing higher measurement accuracy, whilst non-invasive eye-tracking are generally safer and easier to use. Non-invasive eye-tracking relies on either illuminating the eyes with infrared light and measuring the intensity of reflected light (IR Oculography) or recording eye-movement with one or more cameras (Video-oculography), in either the visible or infrared spectrum [Mishra and Bhattacharyya, 2018]. The author notes that inexpensive eye-tracking via built-in cameras in mobile devices such as tablets is becoming increasingly possible. This has the potential of capturing eye-tracking data from a large population of online users without requiring a special

setting[1]. Mishra emphasizes that this contrasts with EEG and brain imaging (such as fMRI), which "require a fairly complex and expensive setup and hence may not be used outside a laboratory" [Mishra and Bhattacharyya, 2018][4].

Mishra and Bhattacharyya [2018] state that while reading text, the relationship between eye movement and cognitive processes can be explained with the *eye-mind hypothesis* (Just and Carpenter 1980). This hypothesis states that "when a subject views a word/object, he or she also processes it cognitively, for approximately the same amount of time he or she fixates on it". In particular, "longer fixations indicate longer processing caused by the word's infrequency and its thematic importance" [Mishra and Bhattacharyya, 2018, 11]. Just and Carpenter state that two central assumptions determine this link [Mishra and Bhattacharyya, 2018, 24]:

- **Immediacy assumption**: A reader attempts to interpret each content word as it is seen, even if he / she must make guesses that may turn out to be incorrect later.
- **The strong eye-mind hypothesis**: The eye remains fixated on a word until its processing is done.

Mishra and Bhattacharyya [2018] notes that gaze patterns and cognitive effort are believed to correlate with the conceptual difficulty of a text. "Linear and uniform-speed gaze movement is observed over texts having simple concepts, and often nonlinear movement with non-uniform speed over more complex concepts [Rayner, 1998]". According to Rayner and Duffy [1986], fixation time is associated with word frequency, verb complexity, and lexical ambiguity, all contributing to lexical complexity. Both Demberg and Keller [2008] and Von der Malsburg and Vasishth [2011] state that complex eye-movement can arise from syntactic complexity, with the latter showing that "complex saccadic patterns (with higher degree of regression) are related to syntactic re-analysis arising from various forms of syntactically complex structures (e.g., garden-path sentence)" [Mishra and Bhattacharyya, 2018, 11].

2.3.2 Electroencephalography (EEG)

Electroencephalography (EEG) is a method that allows recording the spontaneous electrical activity of the brain over time. It is generally noninvasive, with an array of electrodes placed on the scalp. EEG detects voltage fluctuations, caused by ionic

[1] I remark that in recent years, several publications have outlined far-reaching privacy implications of ubiquitous eye-tracking, in particular regarding sensitive and identifying personal information that may be inferred from collected data (see e.g. Liebling and Preibusch [2014]). Potential mitigations are discussed in e.g. Steil et al. [2019].

current inside brain neurons. (see e.g. Niedermeyer and da Silva [2004]). For various clinical tasks (e.g. tumor diagnosis), it has been superseded by high-resolution anatomical imaging techniques, the one discussed in this thesis being functional magnetic resonance imaging (fMRI). However, despite its lower spatial resolution, it continues to be in use, due to its high, millisecond-range *temporal* resolution, among other, more practical considerations [Wikipedia, 2020]. Compared with eye-tracking, EEG measures physiological processes more directly Hollenstein et al. [2019] note it also measures language processing and that invidiual electrodes are related activity to specific brain regions. As such, it also captures spatial aspects of the brain.

External stimuli trigger what EEG registers as *event-related potentials* (ERPs), yielding various components that give insight into different cognitive processes. A prominent component is P300, referring to a positive deflection of the signal occurring approximately 300 ms after a stimulus is presented. It is considered to measure allocation of resources when directing attention, as well as speed of cognitive processing [Dietrich and Kanso, 2010]. In the 1980s, the N400 component was discovered, which peaks at around 400 ms and is related to meaningful stimuli (including words) and has proven fruitful in studying almost the entire range of human language processing [Kutas and Federmeier, 2011]. Hauk and Pulvermüller [2004] also provide evidence that word frequencies modulate early electrophysiological brain responses, suggesting that after presentation of written word stimuli, lexical access occurs after less than 200 ms. When listening to natural speech, electrophysiological responses are reflective of semantic differences between individual words and the previous contexts and are observed after only a brief delay [Broderick et al., 2018].

How EEG signals are mapped to the the stimulus text depends on the stimulus medium. For an auditive stimulus, signals can be time-locked to the onset time of each spoken word in the audio file, as done in Huth et al. [2016]. In the case of self-paced reading of text, a combination with eye-tracking is possible, as seen in Hollenstein et al. [2018] and Hollenstein et al. [2020]. In this case, the EEG signal is time-locked to the onset of word fixations.

EEG can be separated into frequency bands using band-pass filtering, obtaining neural oscillations (brainwaves) at various frequency bands. In Hollenstein et al. [2019] bands range from *theta* (4 - 8 Hz) to *gamma* (30.5 - 49.5 Hz), relating to different cognitive functions. However, in line with Hollenstein et al. [2019], the present thesis works with the raw signal, where the activity of electrodes is encoded by a vector matching their number. I note that further study into the predictability of specific frequency bands may offer additional insight into specific properties of embeddings. Contrary to eye-tracking, EEG yields two-dimensional spatial data.

As such, prediction errors can be related to signals measured by specific electrodes on the scalp Hollenstein et al. [2019], which are associated with activity in particular brain regions. In the present evaluation, no such analysis is possible, as an aggregated, dimensionality-reduced representation is leveraged (see Chapter 6).

2.3.3 Functional Magnetic Resonance Imaging (fMRI)

Functional magnetic resonance imaging (fMRI) measures and maps cognitive activity by detecting changes that are related to blood flow in the brain. Like EEG, fMRI is a spatial approach, however as a brain imaging method, it generates a three-dimensional map of the entire brain with each scan. It segments the brain into millimeter-sized cube-shaped sections, which are called **voxels**, the volumetric equivalent of pixels [Hollenstein et al., 2019]. As such, it is the most direct measurement of cognitive activity presently evaluated, allowing to relate stimuli directly to specific areas in the brain.

The temporal resolution is significantly lower than for eye-tracking or EEG, with a scan taking approximately two seconds. Given the focus of this thesis on sentences, only continuous stimuli are considered, as they are present when listening to spoken text or performing self-paced natural reading [Hollenstein et al., 2019]. In this case, a scan encompasses cognitive processing data from multiple words. Another important difference is the importance of the body's **haemodynamic response** for this method of measuring cognitive signals. Upon neural activity, it leads to a highly transient, localized increase in oxygen delivery, blood flow and oxyhemoglobin. This non-instantaneous physiological response causes the response of the brain to a stimulus being delayed by several seconds and then decaying slowly again over a number of seconds [Miezin et al., 2000]. The onset delay must be compensated in post-processing. Hollenstein et al. [2019, 5] further note: "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."

The explanatory power of fMRI for exploring cognitive semantics is substantiated by the landmark study carried out by Huth et al. [2016]. The authors investigate regions of the cerebral cortex which are described as the *semantic system*. Using fMRI data obtained from subjects listening to hours of narrated stories, they find that the semantic system show a high degree of fine-grained organisation and exhibits patterns which are consistent across individuals. These commonalities allow the authors to curate a semantic atlas, mapping functional areas in the cortex that are semantically selective.

As with EEG, this thesis draws on raw fMRI data (voxel arrays), available either already aggregated on the sentence level or as individual scans to be averaged sentence-wise. This data is then dimensionality-reduced in order to obtain a sentence representation that makes evaluation possible, given the current sparsity of EEG data on the sentence level (see Chapter 6). As such, spatial properties cannot be considered. In future work, it could prove fruitful to explore relationships between word and sentence embeddings and areas of the semantic system. This may be feasible by aggregating information across regions of interested (ROI) defined in reference to the previously mentioned atlas [2].

[2] An example of data ROI-aggregated by the authors themselves is the basis of Brennan et al. [2016], which could not be considered in this thesis due to there being insufficient data on the sentence level.

3 Related work

Following, I first present examples of previous work leveraging recordings of cognitive data in the context of NLP. I then consider extrinsic and intrinsic methods of sentence evaluation and conclude with a summary of the intrinsic cognitive evaluation by [Hollenstein et al., 2019], on which this thesis build.

3.1 Cognitive signals, neural networks and semantics

Wehbe et al. [2014b] note that the process of accessing word meanings from memory and integrating with context mirror how language-processing neural networks attempt to predict incoming words. They compare the cognitive representation with latent layer of a network, summarizing pertinent prior information. The output probability of an input word in a neural network then mirrors the effort expended by the brain, which according to a commonly held hypothesis, is inversely proportional to its predictability [Frank et al., 2013]. The authors let subjects read a chapter from *Harry Potter and the Sorcerer's Stone* and extract, among other things, syntactic and semantic features, as well as discourse level features. Like Huth et al. [2016], they differentiate areas of the brain based on types of represented information.

[Pereira et al., 2018] develop a decoder for inferring semantics of words, phrases and sentences from brain activation patterns of subjects as they are reading natural text. The decoder receives as input brain activation patterns, recorded for each participant when reading individual words, and predicts a semantic vector. In Hollenstein et al. [2019] as well as the present work, this relationship is reversed, with embedding vectors predicting cognitive data in the context of a neural regression. Nonetheless, the intuition is shared that "variation in each dimension of the semantic space would correspond to variation in the patterns of activation" and that the relationship between the two can be learned [Pereira et al., 2018][2]. The authors show that a decoder trained with limited word-level data can decode sentence-level semantics, represented as an average of word-level vectors.

3.2 Embedding evaluation

Hollenstein et al. [2019] broadly distinguish between methods *extrinsic* and *intrinsic* evaluation of embeddings. Extrinsic evaluation measures the quality of embeddings by their performance in downstream tasks, while intrinsic evaluation is concerned with inherent properties of embeddings.

3.2.1 Extrinsic evaluation

A large-scale evaluation of extrinsic performance of sentence embeddings can be found in Perone et al. [2018]. The authors evaluate the main performance measures in the respective task categories. The datasets are provided by the **SentEval** framework introduced by Conneau and Kiela [2018] in an effort to curate a standardized test battery for evaluating sentence embeddings in downstream tasks. Evaluated tasks categories and measures are:

- **Classification tasks** (*Accuracy*) such as sentiment (customer reviews, movie reviews and Stanford sentiment analysis datasets, see Socher et al. [2013] and question answering such as MPQA [Wang and Manning, 2012] and TREC [Li and Roth, 2002].
- **Semantic relatedness tasks** (*Pearson correlation*) such as image-caption retrieval (COCO, Lin et al. [2014]), paraphrase detection (MRPC, Dolan et al. [2004]) and semantic text similarity (STS, e.g. Agirre et al. [2016]).
- Two **information retrieval tasks** (Recall at 1, 5 and 1, R @ x) for caption retrieval and image retrieval on the COCO dataset.

A full overview of the first two categories, can be found in Appendix B. The authors find three sentence embedding types to perform best in the context of classification tasks: The transformer variant of the **Universal Sentence Encoder** [Cer et al., 2018] (USE Transformer), **InferSent** [Conneau et al., 2017], and an average BoW baseline representation of the largest pretrained **ELMo** model [Peters et al., 2018]. In the context of semantic relatedness and textual similarity tasks, USE Transformer performs best in nearly all tasks.

3.2.2 Intrinsic evaluation

3.2.2.1 Linguistic evaluation

Perone et al. [2018] also evaluate ten *linguistic probing tasks*, an extension to **SentEval** introduced by Conneau et al. [2018]. The authors note that they comprise of classification tasks capturing simple linguistic properties of a sentence. Every tasks captures a distinct property, such as the subject or direct object number of the main clause, the tense of the main verb (past or present) and as well as whether a frequently occurring constituent sequence is present. Other tasks detect *anomalies* such as inversion of two words in a sentence (bigram shift) and random replacements of nouns and verbs by other nouns and verbs respectively (semantic odd man out). A full overview of the tasks can be found in Appendix B.

ELMo performs best on seven of the ten linguistic probing tasks. This shows that intrinsic evaluation results partially match extrinsic results, although with limited overlap. The transformer variant of USE, which is highly competitive in the extrinsic evaluation, shows middling to poor performance in most linguistic probing tasks. On the other hand, the good performance of ELMo in both evaluation speaks to some correlation between the approaches, suggesting that the probing tasks may suffer from coverage problems instead, not reflecting the full range of encoded information benefitting downstream tasks.

For both extrinsic and intrinsic evaluation, no approach performed consistently well across all tasks, and Perone et al. [2018] argue that the results are mostly influenced by the similarity of the training task of the approaches and the downstream tasks. Perone et al. [2018] hypothesize that the encoders are too narrowly scoped in what information they model. They state that language models allow to capture context and meaning and note that along with sentence encoding techniques used by the more advanced approaches, this leads to significant improvement of sentence encoding performance. The authors conclude that context-dependent semantics and linguistic features can be better captured by incorporating language models as well as multi-layered representations, as found in ELMo. I note that this would be confirmed shortly after with the innovation of BERT [Devlin et al., 2018].

The evaluation by [Perone et al., 2018] predates BERT and more recent developments with substantial performance improvements. However given that no prior large-scale cognitive sentence evaluation exists, I argue that is useful to assess a cross-section of advances in recent years prior to assessing newer approaches. The results serve as the basis to assess the correlation between predictions of cognitive

signals and extrinsic performance.

3.2.2.2 Cognitive evaluation

Hollenstein et al. [2019] give a brief summary of previous work relating to cognitive embedding evaluation. The field was pioneered by Mitchell et al. [2008], who predicted patterns of neural activation obtained through fMRI using embeddings. Subjects are exposed to 60 word stimuli in isolation (nouns). Various later work builds on this dataset, such as Abnar et al. [2017], who evaluate a variety of embeddings. Søgaard [2016] presents a preliminary evaluation of embeddings with respect to continuous text stimuli obtained via eye-tracking and fMRI. Beinborn et al. [2019] evaluate ELMo with respect to its ability to predict brain responses across multiple datasets. In the context of EEG, e.g. Schwartz and Mitchell [2019] find that aspects of measured ERP components can be predicted by means of neural networks and word embeddings.

Hollenstein et al. [2019] motivate their approach by stating that embeddings which are tailored toward specific tasks lead to good performance within narrow parameters, but may not accurately reflect the semantics of words from a linguistic or cognitive perspective. Existing approaches to linguistic evaluation on the other hand, focus on linguistic aspects in isolation and is based on conscious human judgement, which is potentially confounded by bias that may arise from both the task and subjective factors. They note that intrinsic and extrinsic evaluation metrics do not clearly correlate, as the former fail to predict performance regarding the latter (e.g Chiu et al. [2016]; Gladkova and Drozd [2016]). Furthermore, they find that published intrinsic and extrinsic evaluation results are rarely tested for statistical significance and do not provide a global score regarding quality. I note that both the unclear correlation, as well as the isolated and somewhat artificial nature of linguistic evaluation are evident from the linguistic probing tests described in the previous section, which separately test simple properties.

Søgaard [2016] argues that human brain activity data recorded while language is processed, is the most accurate mental lexical representation available. Hollenstein et al. [2019] summarise further evidence from cognitive neuroscience as follows: Murphy et al. [2018] show that words activate neurons in various brain regions depending on their type. Huth et al. [2016] found semantic maps of the distribution of words throughout the human cortex for subjects listening to stories while their brains are scanned using fMRI. Furthermore, [Frank and Willems, 2017] find that the predicatability and semantic similarity of words are reflected in distinct brain activity

patterns observed during language comprehension. In particular, semantic distance is found to have distinguishable neural effects. Hollenstein et al. [2019] conclude that these findings substantiate the theory that brain activity recordings reflect lexical semantics and is therefore a suitable basis to determine the quality of word embeddings.

To mitigate issues of previous evaluation approaches, Hollenstein et al. [2019] focus on what [Bakarov, 2018] describes as *intrinsic subconscious evaluation*. In this method, word embeddings are evaluated by way of their relationship to lexical representations of words in the human brain, which are recorded whilst subject passively understand language. The authors present the first multi-modal evaluation of English word embedding vectors leveraging cognitive lexical semantics. Hollenstein et al. [2019] note that the size of the few publicly available cognitive datasets is insufficient to be used as training data for advanced machine learning methods. Particularly, distance correlation between representations has insufficient statistical power to compare embedding types, as cognitive data contains highly noisy signals [Frank and Willems, 2017].

Instead, word embeddings (functioning as independent variable) are evaluated by how well they can predict cognitive language processing data (functioning as dependent variable) in a neural, non-linear regression. The authors build on Søgaard [2016]'s theory of a task-independent approach to evaluating embeddings by leveraging cognitive lexical semantics. [Hollenstein et al., 2019] state the three core principles of the CogniVal framework, as established by [Mishra and Bhattacharyya, 2018] and maintained in this work:

- **Multi-modality**: Evaluating against multiple modalities of recording human signals to compensate for the noisiness of the data.
- **Diversity** within modalities: Evaluate against multiple datasets within a single modality to ensure the number of samples is as large as possible.
- **Correlation** of results should be evident across modalities and between datasets of the same modality.

The authors evaluate six word embeddings, such as GloVe [Pennington et al., 2014], fastText [Joulin et al., 2017], which captures morphology as well as rare word and unseen compounds in its subword variant, to the more complex and recent developments ELMo [Peters et al., 2018] and BERT [Devlin et al., 2018], producing contextual, bidirectional word representations that relate a word to the sequence (typically a sentence) it occurs in. For the latter, Hollenstein et al. [2019] only leverage the context-insensitive representations of the encoder given that words are evaluated in

isolation. I describe all of the stated approaches except word2vec in more detail in Chapter 5.

Features are aggregated on the word type level (i.e. across all occurrences of a type) and scaled between 0 and 1. For EEG and eye-tracking, the representations are averaged among subjects prior to scaling. The number of features is determined by the number of electrodes in the headcap and the number of recorded eye-tracking features, respectively. fMRI data differs in that an fMRI scan is a very high-dimensional array of voxels which the authors reduce through random sampling of. Furthermore, signals cannot simply be averaged among subjects due to difference in brain morphology, size, etc., thus, each participating subject constitutes a single hypothesis.

The authors test each hypothesis for statistical significance to assess consistency and in order to advance toward a global quality metric which can be combined with other modes of evaluation. The tested hypothesis is that the predictive performance of an embedding in relation to a cognitive data source, is significantly different from the performance of a randomly generated baseline. I adapt the statistical testing approach and describe it in detail in Chapter 6.

Embeddings are tested against seven eye-tracking datatsets, as well as four EEG and fMRI datasets respectively. The datasets encompass varying text material, stimuli types and experimental parameters. I discuss a subset of these sources in Chapter 4, which are suitable in the context of the present evaluation.

Hollenstein et al. [2019] findings show that more recent developments such as ELMo and fastText embeddings with subwords achieve the best overall prediction results across datasets and modalities. With respect to eye-tracking, they observe that general eye-tracking features accounting for the entire reading process of a word appear to be the most easily to predict. For fMRI, results improve with an increasing number of voxels randomly sampled from the scans, with the final sample size chosen being 1000. For EEG, the middle central electrodes see the most accurate predictions, which are known to register activity in the Perisylvian cortex, a brain area associated with language-related processing [Catani et al., 2005]. The authors also find that most embedding types significantly outperform the randomly generated baselines across a wide range of cognitive features. Furthermore, they observe a strong correlation between the three modalities eye-tracking, EEG and fMRI, implying that word embeddings predict actual brain acvitiy and not pre-processing artifacts related to individual modalities. This is substantiated by clear correlation between datasets within a modality. Finally, the authors also find correlation between the regression results and the reported performance of the embeddings in two downstream tasks, question answering and named entity recognition.

From these results, Hollenstein et al. [2019] infer that their approach may not only serve as a means of evaluation but also inform the selection of adequate embeddings for a task, which does not appear to apply to other intrinsic evaluation methods. The authors emphasize that CogniVal can be effectively combined with other intrinsic and extrinsic embedding frameworks. They conclude that for embeddings to accurately encode word semantics, it is advisable that they reflect mental lexical representations.

As stated, I base the present assessment of sentence embeddings on this work.

4 Cognitive data sources

Following, I introduce the cognitive datasets selected for the present evaluation across each of the previously introduced modalities. For each dataset, two example sentences are shown to illustrate the text material. I also describe any preparatory steps applied to datasets in their entirety. Feature-transforming pre-processing steps are applied on-the-fly during evaluation to avoid unintended inference of unseen data points. These steps are described in Chapter 6.

Only a subset of datasets evaluated in Hollenstein et al. [2019] are suitable for evaluation; I exclude sources with very sparse data on the sentence-level and sources restricted to the word-level. For eye-tracking, a subset of most commonly occurring features across sources was selected, with several capturing the entire reading process. Various other features present in some sources, e.g. statistics for words fixated only once, are omitted, given their unclear potential and the considerable computational cost of parameter search and significance testing for these datasets, which are the largest across all modalities.

4.1 Eye-Tracking

4.1.1 ZuCo

(4.1) The director, with his fake backdrops and stately pacing, never settles on a consistent tone.

(4.2) He flew a P-38 Lightning in the North African campaign of November 1942.

The Zurich Cognitive Language Processing Corpus (ZuCo) is a combined Eye-Tracking and EEG dataset curated by Hollenstein et al. [2018].

The subjects are healthy adults who read isolated natural English sentences over the span of four to six hours. The curation of the resource is in line with the long-term goal to replace manual annotation with physiological activity data. As subjects read sentences, opinions and sentiments are evoked, which find expression in their brain

activity. The author's hypothesis is that the recorded signal should be decodable with machine learning techniques to obtain the this information, either complementing or replacing manual annotation efforts. The study comprises three tasks, two normal reading paradigms with different text types and a task-specific paradigm, which required active subject participation to solve a language comprehension exercise. Only the normal reading paradigms are considered in [Hollenstein et al., 2019], which is continued in this thesis. The text material is labelled with relation types and contains sentences from the Stanford Sentiment Treebank (SST, Socher et al. [2013]) on one hand, which were extracted from movie reviews. The second portion consist of biographical sentences obtained from paragraphs about notable people from the Wikipedia relation extraction corpus [Culotta et al., 2006]. The authors emphasized naturalistic reading, with the full sentence presented at once. This allowed subjects to read at their desired speed and to freely choose order and duration of word fixation. The setting contrasts with word-by-word presentation which – the authors emphasize – does not reflect normal reading.

400 single sentences were selected from the SST, which have been manually annotated to be neutral (123), negative (137) and positive (14). From the Wikipedia relation extraction dataset, 300 sentences containing a semantic relation were randomly selected for the normal reading task.

The following eye-tracking features are considered in this thesis: Number of fixations on a word (nFix), "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), first fixation duration (FFD), the duration of the first fixation on the prevailing word, [and] total reading time (TRT), the sum of all fixation durations on the current word, including regressions" [Hollenstein et al., 2018][7]. I note that neither sentiment nor relations are considered here and sentences from both corpora are combined into one source in order to maximize available data in the context of this evaluation.

4.1.2 Dundee

(4.3) Ofsted has also found problems with the quality of supply staff, saying that they often missed out on specialist training and were not as good as permanent teachers.

(4.4) Certainly, growth slows down but, unlike other countries, there is no significant contraction in economic activity.

Barrett et al. [2015] state that at the time of writing, the Dundee Corpus [Kennedy

et al., 2003] was the largest existing eye-movement corpora and a significant resource for studying through eye movements how language is processed in the context of psycholinguistics. The authors note that the dataset enabled, among other things, the study of differences in processing difficulty of words relating to syntax and semantic aspect.

The native English-speaking subjects read 20 newspaper articles, with the English portion of the corpus compiled from articles of *The Independent*. As with ZuCo, order and duration of fixation is available on the word level and subjects read five lines of running text per viewing.

As with ZuCo, I select the number of fixations (nFix), first fixation duration (FFD) and total fixation duration (total reading time, TFD/TRT). Instead of the sum of all fixations (gaze duration), Dundee provides the mean fixation duration (MFD), which I use as substitute. Additionally, I consider fixation probability (FP), which is not available for the the other datasets.

4.1.3 GECO

(4.5) She's the mater's factotum, companion, Jack of all trades!

(4.6) We were detained under suspicion by the hospital porter, until Cynthia appeared to vouch for us, looking very cool and sweet in her long white.

The Ghent Eye-Tracking Corpus (GECO) is a bilingual resource curated by [Cop et al., 2017]. It stands out in that participants read an entire novel, compared to short newspaper articles as for Dundee or isolated sentences as in the case of ZuCo. It has a similar token count as Dundee, but contains twice as many sentences (see below). For the English portion, the subjects are English monolingual undergraduates from the University of Southampton.

The authors note that additional cognitive processes such as sentence integration take place when reading longer texts or narratives, which typically cannot be observed when reading isolated sentences. Subjects read *The Mysterious Affair at Styles* by Agatha Christie. As with Dundee, only the monolingual (English) paradigm of the study is considered. Subjects viewed the novel on a screen in paragraphs, which they could read at their desired speed and advance to the next paragraph at the press of a button. Approximately a third of the words are skipped by participants, which the authors find to be in line with other eye-tracking research.

I select the same features for this corpus as for ZuCo. Among these features, the

authors note that first fixation duration most approximated a normal distribution and that total reading time is more skewed than both first fixation duration and gaze duration, with all measures being right-skewed to some extent.

4.1.4 Dataset specifics and preparation

	Dundee	**GECO**	**ZuCo**
Tokens	52,524	57,237	14,071
Types	9,560	6,002	4,589
Sentences	2,366	5,073	700
Subjects	10	14	12
Medium	text	text	text
Isolated sent.	no	no	yes
Reading material	20 news articles	novel	Wikipedia, Movie reviews

Table 1: Descriptive statistics of eye-tracking datasets, excluding punctuation on the word level. For multilingual datasets, only the English portion is considered. For ZuCo, only sentences from paradigms relevant to the present evaluation are considered. Note that values reflect counts after data preparation and may deviate from statistics reported by the respective authors.

I average word-level features per sentence, among all subjects for whom signals were recorded for that particular sentence. This leads to variable robustness depending on the sentence, but maximizes the available data for subsequent model training. Each of the selected features constitutes a hypothesis and is evaluated individually. Any null values encountered are filled with zeros. For ZuCo, correctly formatted sentence strings are obtained by simply reconcatenating token strings, which are available with trailing punctuation. For Dundee, I use the version provided by [Barrett et al., 2015], with eye-tracking features extracted from the original corpus. Punctuation is represented as separate tokens, thus I reconstruct sentences using a simple heuristic detokenization routine[1]. For GECO, sentence strings are provided in a separate file and are mapped to word-level features accordingly. Inexplicably, more than 10,000 word IDs (scattered seemingly random throughout sentences) do not map to sentence IDs. I use interpolation to fill in the gaps, as the sentence ID of a word followed and preceded by words with identical sentence ID can be inferred readily. Notably, this potentially leads to errors at sentence boundaries.

As eye-tracking features are scalars, the token-level sequence of features can be used without dimensionality issues, as the vector size is bounded by the maximum

[1]The routine has been adapted from `https://stackoverflow.com/a/59618856`

sequence length[2]. This contrasts with dimensionality reduced vectors obtained for the high-dimensional EEG and fMRI data, in that it allows to fully retain sequence information. The output format is a matrix of zero-padded sequences of eye-tracking measures per sentence, with one dataset per eye-tracking feature.

4.2 Electroencephalography (EEG)

4.2.1 ZuCo & ZuCo 2

(4.7) The Perrys have four children. (ZuCo 2)

(4.8) She was a researcher and reporter at Thames Television [1980 to 1983]. (ZuCo 2)

I also consider the EEG portion of both ZuCo and the similarly structured extension ZuCo 2 [Hollenstein et al., 2020]. EEG signals are timelocked to the onset of word fixations recorded by the eye-tracking apparatus. This allows to identify word boundaries and extract word-level EEG signals accordingly. As with ZuCo, ZuCo 2 comprises of sentences from the English Wikipedia, sampled from a corpus provided by Culotta et al. [2006]. The subjects are healthy English native speakers. The authors note that there are more fixations for normal reading tasks, specifically for the entire sentence, whilst for task-specific reading, fixations stop after arriving at the salient words. For ZuCo 2, 50 sentences of normal reading and task-specific reading are read in alternating blocks. Sentences are read during a single session with a duration of between 100 and 180 minutes. As with ZuCo, only normal reading sentences are considered. In the context of this evaluation and regarding the considered experimental data, there are no central differences between the datasets other than the reading material, structure of sessions and number of data points.

I note that due to time-constraints and given the significantly smaller size of the the ZuCo 2 dataset compared with other eye-tracking sources, only the EEG portion is considered in this thesis.

[2] 93 tokens including punctuation for Dundee and 69 tokens for GECO, and 65 tokens for ZuCo respectively, excluding punctuation

4.2.2 Natural Speech

(4.9) When the boy came back the old man was asleep in the chair and the sun was down.

(4.10) The second was at seventy five and the third and fourth were down in the blue water at one hundred and one hundred and twenty five fathoms.

Broderick et al. [2018] present a dataset consisting of subjects listening to narrative speech played back from an audiobook. Each subject participated in 20 trials with a length of approximately 180s, where they listened to what the authors describe without further specification as a popular mid-20th century American work of fiction with understated writing, which is read by a single male American speaker at an average rate of 210 words per minute. Hollenstein et al. [2019] find the data of only 14 of a total of 19 subjects to be of sufficient quality, and this selection is retained in the present evaluation.

4.2.3 Dataset specifics and preparation

	Natural Speech	**ZuCo**	**ZuCo 2**
Tokens	11,416	14,108	6,889
Types	1,736	4,587	2,491
Sentences	695	700	344
Subjects	14	12	18
Medium	speech	text	text
Isolated sents.	no	yes	yes
Reading material	novel	Wikipedia, movie reviews	Wikipedia

Table 2: Descriptive statistics of EEG datasets, excluding punctuation on the word level. For both ZuCo datasets, only sentences from paradigms relevant to the present evaluation are considered. Note that values reflect counts after data preparation and may deviate from statistics reported by the respective authors.

EEG signals of all fixations of a word are averaged. To obtain the sentence representation, the resulting word-level vectors are concatenated and zero-padded, as with eye-tracking, which was found to outperform averaging of word vectors during initial evaluations. Given data sparsity and the high resulting dimensionality of 11050 (Natural Speech), 6825 (ZuCo) and 5565 (ZuCo 2) respectively, dimensionality reduction is necessary, and performed on the fly during evaluation. This is informed by the notion that given the indirect approach of measurement and standardized

geometry of EEG caps not optimized for studying language processing, signal redundancy is to be expected in raw data. Furthermore, Hollenstein et al. [2020] note that EEG data is noisy and can contain bad channels. As with eye-tracking, vectors are averaged for all subjects for which EEG data was recorded for a particular sentence.

For both ZuCo datasets, raw EEG data has a dimensionality of 105, corresponding to the number of electrodes in the headcap. Words with missing fixations are assigned the zero vector. A total of 700 sentences are obtained from ZuCo, 400 from the sentiment-focused and 300 from the relation-focused task. 345 sentences from the normal reading task are obtained from ZuCo 2.0, with an overlap of 36 sentences with ZuCo, which is retained. Analogous to the ZuCo eye-tracking data, sentence strings are restored by simply concatenating tokens.

For Natural Speech, the number of electrodes and resulting dimensionality is 130. The EEG signal is time-locked to the speech stimulus, and both onset and offset of a spoken token are recorded. Thus, only one vector is available per token, contrary to both ZuCo corpora, where multiple fixations of a word are possible due to the natural reading setting. The corpus has been preprocessed by Hollenstein et al. [2018] using the reported methods the authors applied to ZuCo and I reuse the preprocessed variant here. As with both ZuCo datasets, the raw EEG representation is used. Sentences are reconstructed from separately stored sentence-boundaries, which are represented as offset times matching certain tokens. Tokens are not cased and punctuation has been discarded. To obtain an approximation of the original sentence string, the first token and any occurrence of the frequently occurring pronoun "I" are capitalized and a period is appended. This is to ensure that parsers and encoder-based embedding approaches such as BERT correctly recognize the sequence as a sentence. Named entities remain uncased. Two short sentences have four occurrences each, the representations of which I average, ultimately obtaining 695 sentences.

4.3 Functional Magnetic Resonance Imaging (fMRI)

4.3.1 Pereira et al. [2018]

(4.11) It's even worse when you are submerged in the frigid waters directly beneath the huge animal. (experiment 2)

(4.12) The team of astronauts floated out together to the exterior of the space shuttle. (experiment 2)

(4.13) A mitten is a kind of glove where the four fingers are covered together. (experiment 3)

(4.14) The small sharp points along the cutting side of the saw are called the teeth. (experiment 3)

Pereira et al. [2018] offer a dataset comprising three experiments, with only the datasets of the last two representing sentences as strings. The subjects are either native speakers or in one case, bilingual with native-like fluency. Only eight of the originally 16 participants participate in experiment 2 and only five in experiment 3, with the latter being a subset of the former. In both experiments, text consists of passages with a length of three to four sentences, covering 48 broad topics such as professions, clothing, birds, skiing, dreams and opera musical instruments. Each topic is represented by three to four passages (such as clarinet, accordion, etc. for instruments) and there is no overlap in topics between the two experiments. Both experiments contain passages in the style of Wikipedia, offering a basic description of the respective concept. A third of the passages in the smaller experiment are represented by narratives (both first- and third-person).

For both experiments, sentences of passages are presented individually, and every passage is viewed three times by each subject. Sentences are shown for 4 seconds, followed with an equally long pause. This enabled the researchers to obtain sentence-wise scans.

4.3.2 Wehbe / Harry Potter

(4.15) Harry took out his wand in case Malfoy leapt in and started at once.

(4.16) What do they think they're doing, keeping a thing like that locked up in a school? said Ron finally.

The dataset curated by Wehbe et al. [2014a] contains continuous fMRI scans of native English speakers, who read chapter 9 of the novel *Harry Potter and the Sorcerer's Stone*[Rowling, 1999]. Notably, all subjects were either familiar with the book series or movie adaptations prior to participation. The authors emphasize the non-artificial character of the text, noting that it exposes the subjects "to the rich lexical and syntactic variety of an authentic text that evokes a natural distribution of the many neural processes involved in diverse, real-world language processing" [Wehbe et al., 2014a][3]. In particular, the authenticity of the material is considered by the authors to be more engaging for the subjects, helping to sustain their attention throughout the experiment.

The text was presented using *Rapid Serial Visual Presentation* (RSVP). With this method, words are presented sequentially in isolation, with a fixed duration of 0.5s. The entire chapter was presented in 45 minutes. Every two seconds, an fMRI image is obtained. Wehbe et al. [2014b] note that changes in the obtained signal are persistent over approximately eight seconds after the onset of neural activity.

4.3.3 Dataset specifics and preparation

	Pereira (large)	**Pereira (small)**	**Wehbe**
Tokens	4,558	3,241	5,102
Types	1,623	1,411	1,348/1,349
Sentences	384	243	398
Subjects	8	5	8
Medium	text	text	text (RSVP)
Isolated sents.	no	no	no
Reading material	Wiki. paragraphs	Wiki. paragraphs	novel chapter

Table 3: Descriptive statistics of fMRI datasets, excluding punctuation on the word level. The two experiments from Pereira are treated as separate datasets. For Wehbe, there are minimal differences between subjects. Note that values reflect counts after data preparation and may deviate from statistics reported by the respective authors.

For all datasets, pre-processing is performed using a modified version of the *brain-lang*[3] toolkit. The raw scans are used without any pre-processing, other than sentence-level aggregation (see below). The output representation is a flat one-dimensional array, as with eye-tracking and EEG. fMRI voxel data has very high dimensionality, ranging from approximately 27,000 to 38,000 voxels for Wehbe and 145,000 to 201,000 for Pereira. Regardless of data sparsity, sampling or dimensionality reduction is thus unavoidable. Hollenstein et al. [2019] randomly samples voxels, with a better performance obtained at 1000 voxels than at smaller sample sizes. Given the significantly reduced dataset size, I considered it necessary to minimize dimensionality whilst ensuring that dimensions of the final vector represent the fMRI signal as efficiently as possible. The procedure is described in Chapter 6. Contrary to EEG and eye-tracking, raw fMRI data cannot be averaged among subjects due to the differences in brain morphology.

For the Pereira dataset, only the second and third experiments contain sentence data and are therefore considered in this study. The sub-datasets are treated separately,

[3] `https://github.com/beinborn/brain-lang`

as the number of subjects differ and a concatenation of data between shared subjects has not lead to an improvement in initial evaluations. A single scan is available per sentence, thus only dimensionality reduction is necessary to obtain a usable representation. `brain-lang` provides a reader for the first experiment, which I modify (and simplify) to read and convert the relevant raw data. The sub-datasets are referred to as *Pereira (large)* and *Pereira (small)* from now on.

The toolkit also provides a reader for the Wehbe dataset. Sentence boundaries are not explicitely given and therefore detected with a simple heuristic the authors tuned to the dataset. I note that there are minor differences in punctuation representation, leading to a deviation in the number of sentences, but not tokens. In relation to this dataset, [Beinborn et al., 2019] discuss the hemodynamic delay, which needs to be compensated to align stimuli and scans. It is approximated with two timesteps, with a timestep corresponding to the fMRI scan duration of 2 seconds in the study carried out by Wehbe et al. [2014a]. During preliminary tests, I obtain slightly lower errors when not applying a delay and thus omit it. Scans are aligned with ranges of tokens read by the subject during the scan. To obtain a sentence representation, I modified the pipeline to average all scans associated with a sentence ID, with scans straddling sentence boundaries factoring into both the current and the next sentence.

4.4 Discussion

The cognitive datasets discussed above not only differ in their modalities but also in the type and variety of the presented material and the methods used for presentation. Regarding content, ZuCo and Pereira mark a middle ground, with either isolated sentences or short paragraphs being shown yet a large variety of concepts being presented. In both instances, extent of sentence integration as discussed by Cop et al. [2017] is either non-existent or likely to be very limited. Between these sources, the movie review portion in ZuCo is arguably the most natural text genre, while Wikipedia-based sentences from either sources typically conform to encyclopedic conventions and are more formulaic whilst comprising a broad range of concepts. More naturalistic material is found in the Dundee corpus, where entire newspaper articles are read, however coming from a single newspaper, variety is likely to be more constrained by editorial conventions and focus. The most natural instances are Natural Speech (listening to audiobook snippets), Wehbe (reading a book chapter) and GECO (reading an entire book), with the caveat that all sentences are now obtained from a single author document, further constraining discourse variety, but

likely offering a broader range of lexical semantics and syntactic phenomena. Furthermore, Wehbe et al. [2014b] note that the read chapter has frequent occurrences of direct speech.

Curiously, presentation methods do not parallel this spectrum. ZuCo allows for self-paced reading of sentences, while Pereira dictates the maximum duration a sentence is viewed, constituting an interference with the natural reading process. Dundee and GECO are also self-paced, whereas Natural Speech trivially fixes the listening speed[4]. Finally, Wehbe relies on rapid serial visual presentation, which is the most artificial method of text consumption found across all datasets considered in this thesis. As previously mentioned, it is also arguably the most problematic in that it is unlikely to capture all sentence-level reading processes, as e.g. regressions to earlier words are not possible.

In summary, the discussed datasets present significant differences between modalities, modes of data preparation, reading material and curation methods. This is compounded by sparse dataset sizes, with the exception of eye-tracking sources Dundee and GECO. In light of this heterogeny, assessing the performance of sentence embeddings with respect to genre, specific syntactic or semantic occurrences and other aspects is very difficult. Such an attempt would require controlling for all potentially confounding variables, which is not possible in a high-level multi-modal evaluation. Because of this, I do not primarily consider aspects of individual datasets and instead follow the core principles proposed by [Mishra and Bhattacharyya, 2018] by mainly considering correlation within and among modalities, i.e. patterns on the macroscopic level. I note that ZuCo presents a special case, given that subjects and text material are held constant across modalities.

[4] I expect however, that creators of professional audiobook recordings test and account for what is considered a comfortable reading speed by the majority of consumers.

5 Sentence embeddings

The present work constitutes a high-level evaluation of the general possibility of predicting sentence-level cognitive signals with sentence embeddings. Given this scope, the architecture and properties of individual word and sentence embeddings can be covered only briefly and much of the theoretical foundation is therefore only touched upon, or omitted. For details, I refer the reader to the cited literature.

I note that I subsequently omit several notable and more recent developments such as *ERNIE 2.0* [Sun et al., 2020]. As this work is, to my knowledge, the first multi-modal cognitive sentence embedding evaluation, my goal is not primarily to benchmark state-of-the-art approaches. Instead, the present evaluation has a more longitudinal character, with selected approaches differing in how the sentence representation is obtained, if contextual information is encoded and whether supervised data is leveraged. In particular, I'm interested whether and how the (proven) differences in downstream evaluations are reflected when predicting cognitive signals. In this context, approaches are of particular interest which perform comparably in downstream applications, yet are conceptually different. Discrepancies in the performance of predicting cognitive signals for such approaches are indicative of different types of features predicting specific cognitive modalities.

5.1 Sentence embeddings

Sentence embeddings constitute an extension of distributional semantic representations such as *GloVe* and *word2vec*, which have been popularized some years ago and significantly advanced the state of the art in natural language processing. These approaches and various developments were confined to word-level semantics and are context-independent in that only a single representation is obtained per word. Later efforts such as *ELMo* allowed to encode sentence-level semantics into word representations, by assigning each token a representation which is a function of the full input sentence [Peters et al., 2018]. Sentence-level representations are typically obtained from word-level approaches by averaging word-level representations, obtaining a

Bag-of-Words (BoW) representation, which does not retain sequence information. In parallel and starting with *Skip-Thought*, sentence encoders were developed, which directly learn a sentence representation without the need of aggregating word-level representations, and explicitly consider the sequence of words. Pereira et al. [2018] note that these representations also allow to predict human similarity judgements of paraphrases on the phrase- and sentence-level.

Following, I discuss a selection of six approaches to obtaining sentence representations, representing a subset of methods evaluated by [Perone et al., 2018]. This allows to relate the present results to the large-scale extrinsic and intrinsic results presented by the authors. In addition, GloVe with a dimensionality of 50 serves to establish a lower bound. I discuss simple and more sophisticated averaging baselines, self- or unsupervised approaches, as well as supervised methods.

5.2 Motivation

5.3 Baselines

The simplest method to obtain a sentence representation is to perform a component-wise arithmetic mean of word-level representations, which has been found to be a strong baseline (see e.g. [Kenter et al., 2016]). In the case of approaches not encoding information about the word context, typically only content words such as (proper) nouns, adjectives, main verbs and sometimes adverbs are considered, as not to dilute the representation with stopword vectors.

5.3.1 GloVe

GloVe (short for Global Vectors) embeddings [Pennington et al., 2014] are generated by performing an aggregation of the global word-word co-occurrences in a corpus. The model thus directly captures global statistics of the corpus, relying on counts of word occurrences in the context of other words (represented by a word-word co-occurrence matrix).

In the present evaluation, this approach only serves to establish a lower bound, hence a low-dimensional variant is used. [Hollenstein et al., 2019] show that it is matched or outperformed by fastText on the word level in most instances (for a superset of the cognitive datasets presently evaluated). For this reason, I omit a

detailed description of this approach.

5.3.2 fastText

fastText embeddings [Joulin et al., 2017] are generated using a shallow linear model, contrasting with non-linear neural network approaches. Originally introduced as an efficient means of text classification, the hidden layer representation is frequently used as a word embedding. Training time is greatly improved using hierarchical softmax [Goodman, 2001]. A bag of n-grams is used to partially capture the local word order, which has found to be efficient and comparable in performance with approaches considering the word order explicitely. Finally, the hashing trick [Weinberger et al., 2009] allows fast n-gram mapping with a low memory footprint [Joulin et al., 2017].

fastText saw an extension to account for morphological variations and moving from CBOW to skipgram as its basis. This is done by learning vectors for character n-grams of words, or *subwords*. It derives from the skipgram model with negative sampling as introduced by Mikolov et al. [2013b]. Context words are predicted independently from each other using binary classification, with random sampling from the dictionary being used to obtain negative examples. Subwords are obtained by representing each word as a bag of character n-gram, with a range of n-gram sizes being considered. An unseen word is thus represented as the sum of the subword representations corresponding to its n-grams [Bojanowski et al., 2017].

5.3.3 Concatenated Power Mean Word Embeddings (Power-Mean)

Power-Mean constitutes a special case in that it is conceived as a baseline from the outset, which seeks to be more competitive than previous averaging baselines. Rücklé et al. [2018] motivate their approach by stating that existing averaging approaches to sentence embeddings are likely to benefit from an increase in dimensionality. The authors propose to increase the information content by concatenating word embedding types capturing differing linguistic aspects (syntactic, semantic, sentiment-related, etc.) They also examine benefits of generalizing the averaging operation to the *power mean* [Hardy et al., 1952], which encompasses many types of mean, including common types such as the arithmetic mean ($p = 1$), the geometric mean ($p = 0$), and the harmonic mean ($p = -1$):

$$\left(\frac{x_1^p + \ldots + y_n^p}{n}\right)^{1/p}; p \in \mathbb{R} \cup \{\pm\infty\} \tag{5.1}$$

For ($p = -\infty$) and ($p = +\infty$), the power mean calculates the minimum and maximum of the input sequence, respectively. In the following equation, $H_p(W)$ represents the component-wise power means of the individual word embedding vectors of a sentence s, for a particular value of p[1]. $\mathbf{s}^{(i)}$ is then the concatenation of K different power means of the sequence, i.e. for an array of different values of p.

$$\mathbf{s}^{(i)} = H_{p1}(\mathbf{W}^{(i)}) \oplus \cdots \oplus H_{pK}(\mathbf{W}^{(i)}) \tag{5.2}$$

This transformation is applied to different word embedding types and the resulting power mean sentence representations $\mathbf{s}^{(i)}$ are, again, concatenated to obtain the final representation:

$$\bigoplus_i \mathbf{s}^{(i)} \tag{5.3}$$

Rücklé et al. [2018] choose four embeddings as the basis for their approach, which they believe to be potentially complementary: *GloVe* trained on Common Crawl data; *word2vec* [Mikolov et al., 2013b] trained on GoogleNews data; *Attract-Repel* [Mrkšić et al., 2017] and *MorphSpecialized* [Vulić et al., 2017]. Due to lack of space, I omit a discussion of word2vec, referring to the cited literature. The remaining embedding types can be briefly summarized as follows: Attract-Repel embeddings rely on injected constraints which are extracted from lexical resources. This specializes word-level vectors with respect to their semantics, in that both mono- and cross-lingual constraints regarding synonymy and antonymy are injected, leading to unified cross-lingual vector spaces. The authors showed these embeddings to perform well on SimLex-999 [Hill et al., 2015], a dataset of word pairs that have been annotated in such a way that words which are *related* but not *similar*, receive a low score (e.g.*book* and *read*). MorphSpecialized embeddings build on the Attract-Repel method, injecting morphological constraints into vector spaces. The authors claim that inflectional and derivational rules implicitely encode semantic information, which is thus leveraged. For *attract* constraints (making embeddings more similar), inflectional morphology is used, which only concerns grammatical agreement and does not influence the word meaning. *Repel* constraints (driving embeddings apart) are represented by derivational antonyms. Due to space constraints, interested readers are referred to the respective literature for further details.

The best performing variant concatenates all four embedding types, with application

[1]Which, in the case of $p = 1$ for the arithmetic mean, is identical to the averaging I apply to the previously discussed baselines.

of the z-norm [LeCun et al., 1998] to account for the fact that embeddings and power means may differ in ranges. For each embedding, power means of $[-\infty, 1, \infty]$ are calculated[2]. The dimensionality of the vectors finally obtained is 3600.

Across nine classification tasks, Rücklé et al. [2018] find this variant to be competitive with InferSent (Conneau et al. [2017], see below) while being less computationally expensive. The evaluation by Perone et al. [2018] largely reproduces these results for the same corpora, adding two further datasets (and omitting others).

5.4 Self- or unsupervised approaches

5.4.1 ELMo

Contrary to previous approaches, ELMo word representations [Peters et al., 2018] are computed under consideration of the entire input sequence. Formally, they are the result of two-layer bidirectional language models with character convolutions, as a linear function of hidden units of the network. In the forward direction, a language model (LM) computes the sequence probability as the product of the probabilities of every token t_k given its previous tokens $(t_1, ..., t_{k-1})$:

$$p(t_1, t_2, ..., t_N) = \prod_{(} k = 1)^N p(t_k | t_1, t_2, ..., t_{k-1}) \tag{5.4}$$

The token representation is context-independent, in that the authors compute a CNN over characters. This representation is passed through several forward LSTMs (long short-term memory networks, Hochreiter and Schmidhuber [1997]) and at every position k, a context-dependent representation is yielded. The output of the top-most layer is fed into a softmax layer, predicting the next token. The backward LM is formulated analogously, except it models the sequence in the reverse direction. The biLM combines the forward and backward LM with a residual connection between LSTM layers, jointly maximizing the the log likelihood of both representations. For every token, $2L + 1$ representations are computed, (L corresponding to the number of LSTM layers), corresponding to a single token layer and two hidden units for each biLSTM layer (one per direction). In downstream tasks, all layers are collapsed into a flat vector. This can be achieved by either a weighted sum of all biLM layers, with weights tuned in the context of a specific task, or by simply

[2] As such, three variants are the special cases minimum, arithmetic mean and maximum, i.e. do not require formulating the power-mean. However, other values are used in the omitted cross-lingual experiments.

selecting the top-most layer. The pretrained embeddings used in this thesis rely on two biLSTM layers, with every LSTM projected into 512 dimensions, leading to a dimensionality of 1024. The context insensitive representation uses 2048 character n-gram convolutional filters and two highway layers, projected to a dimensionality of 512. As such, the latter can capture unseen words, like fastText. In their evaluation, the authors observe that ELMo disambiguates word meanings (mainly the top layer) in relation to their respective context and also captures basic syntax (mainly the bottom layer).

Given the scope of the present evaluation, I seek to evaluate embeddings in their generic form. Hence, in line with Perone et al. [2018], the three layers of the models provided by the authors are concatenated on a per-token basis to form a 3072-dimensional token representation, which is then averaged.

5.4.2 Skip-Thought

Skip-Thought [Kiros et al., 2015] is an early notable instance of a dedicated sentence encoding approach. It offers a unsupervised alternative to previous approaches, not relying on a specific task or expensive inference at test time. The model applies the skip-gram approach of *word2vec* [Mikolov et al., 2013a], substituting words with sentences, and thus predicting the sentences occurring in the context of a sentence. It is an encoder-decoder model, i.e. words are mapped to a sentence vector by an encoder, and the decoder generates the surrounding sentences. Both encoder and decoder are recurrent neural networks (RNN), with the encoder using gated recrurrent unit activations (GRU) [Chung et al., 2014], and the decoder relying on a conditional GRU. When processing a sentence, the encoder produces a hidden state at each time step, representing the sequence as far as the respective word in the sentence. It follows that the final hidden state represents the full sentence. The decoder is a neural language model, conditioned on the output of the encoder. It performs a similar computation as the encoder, however the update gate, reset gate and hidden state are biased by the sentence vector. Two decoders with separate parameters are used, to predict the previous and subsequent sentences respectively (which constitute the context of the encoded sentence). The vocabulary matrix is shared, which is used to compute a distribution over words. Figure 1 illustrates an example of a sentence triplet, i.e. an encoded sentence and surrounding context to be predicted. The authors note the similarity with neural machine translation approaches at the time. To improve generalizations to other corpora, they learn a linear mapping from word2vec's [Mikolov et al., 2013b] vector space to the vocabulary space of the encoder, for all words shared by both vocabularies. Using this

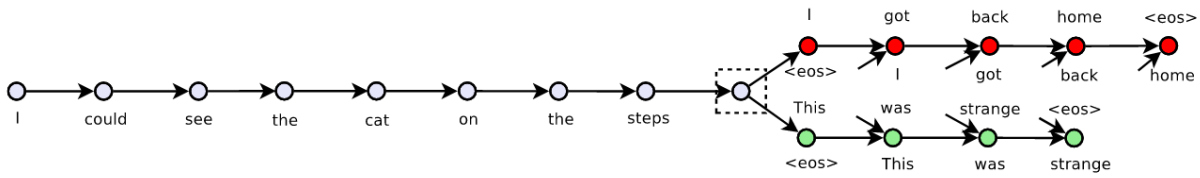


Figure 1: Given a contiguous sequence of sentences (s_{i-1}, s_i, s_{i+1}) the sentence s_i is encoded, while the decoder attempts to reconstruct the previous and next sentences s_{i-1} and s_{i+1} respectively. Here, the sentence sequence is: *I got back home. I could see the cat on the steps. This was strange.* (encoded sentence underlined). Source: Kiros et al. [2015]

learnt mapping, all word vectors in a word2vec model can be transformed into a vector in Skip-Thought's encoder embedding space, extending its vocabulary.

The authors evaluate both an uni-directional and a bi-directional variant, the latter using two encoders, processing the sentences in the forward and reverse order respectively. Both approaches share the same output dimensionality of 2400, with 1200 dimensions allocated per direction for the bi-directional variant. They find that concatenating the uni- and bi-directional output yields the best results, indicating that the approaches are complementary. This representation results in a dimensionality of 4800 and it is the approach considered in the present evaluation.

Rücklé et al. [2018] note the computational cost of Skip-Thought, which is one the primary reasons for later developments that saw improved running time at training and/or testing time. Some of these approaches are described in the following sections. I note that I evaluate the original Skip-Thought model, not the later Skip-Thought-LN variant, which has been found to perform better in various (but not all) tasks (see Conneau et al. [2017]).

5.4.3 BERT

BERT [Devlin et al., 2018] builds on the method of language model pre-training, which the authors note to be effective in improving a large number of natural language processing tasks. It is distinct from the previously discussed ELMo, which relies on a feature-based approach and leverages task-specific architectures, in which datasets are augmented with the pre-trained representations, representing additional features. BERT instead relies on fine-tuning and allows for bidirectionality in language model learning by employing a *masked language model* (MLM) pre-training objective, modelled on the Cloze task [Taylor, 1953].

The presented framework is divided into two steps, unsupervised pre-training and supervised fine-tuning, which are illstrated in Figure 2. **Pre-training** consists of two

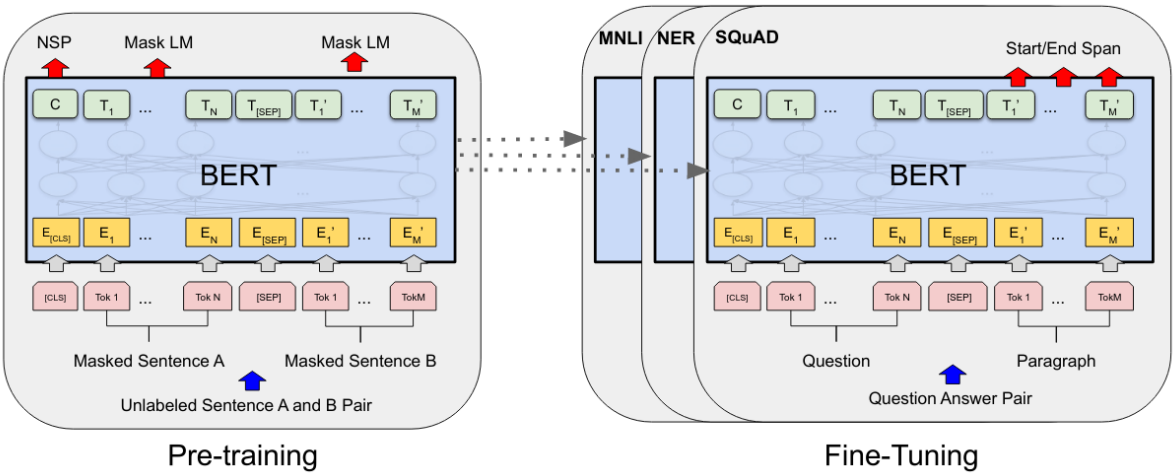


Figure 2: Pre-training and fine-tuning of the BERT architecture. During pre-training, an unlabeled sentence pair is encoded and leveraged in the next sentence prediction (NSP) and masked LM (MLM) tasks. During fine-tuning, previously learnt parameters are adjusted with labeled data from the downstream task. Source: Devlin et al. [2018]

unsupervised tasks, the *masked language model* (MLM) and *next sentence prediction* (NSP). In the MLM task, some percentage of the tokens are randomly masked in the input, with the objective of predicting the ID of the masked word given its context. It is this masking that allows bidirectional conditioning, as it prevents the word from observing itself in other predictions. This objective allows to capture both the left and right context of a word. As in a standard language models, the hidden states associated with the mask tokens serve as the input to a softmax layer over the vocabulary. The authors note that many downstream tasks require that the relationship between two sentences is modeled, which the language model does not capture directly. Examples are Question Answering (QA) and Natural Language Inference (NLI). NSP accounts for this requirement, which can be learned from an arbitrary monolingual corpus. During training, the second sentence is the true next sentence in half of the cases, and randomly chosen otherwise.

During **fine-tuning**, the model is initialized with the parameters previously learnt, which are then tuned, leveraging labeled data from downstream tasks. The authors find the difference between the pre-trained and tuned architecture to be minimal, irrespective of the task type. As with ELMo, this work relies on generic pre-trained BERT embeddings. As such, I do not consider the fine-tuning aspect in further detail.

Devlin et al. [2018] use a multi-layer bidirectional transformer encoder with multi-head (self-)attention, which is almost identical to the approached introduced by Vaswani et al. [2017]. This encoder-decoder (transduction) model relies solely on the

attention mechanism [Bahdanau et al., 2014] to learn global dependencies between input and output. As such, it can be better parallelized and obtain competitive results with faster training times than approaches relying on recurrent or convolutional layers. The attention mechanism allows modeling of dependencies regardless of their distance in input or output sequences. Contrary to previous attention-based transformer uses a constant number of operations to relate signals from pairs of arbitrary input and output positions. Simultaneously, it relies on multi-head attention, allowing it to efficiently learn dependencies between distant positions. I omit a further discussion of the transformer architecture and the attention mechanism for reasons of space and instead refer the reader to the cited literature.

[Devlin et al., 2018] train two model sizes, *Base* and *Large*, where the number of hidden layers are 12 and 24, the number of self-attention heads are 12 and 16 and the hidden-layer sizes are 768 and 1024, respectively. The input is represented by WordPiece tokens [Wu et al., 2016], which attempt to find a balance between vocabulary size and out-of-vocabulary words using a data-driven approach. The authors conclude that pre-trained representations can in many cases supersede task-specific architectures that require extensive engineering. They find that such architectures are often outperformed on both token- and sentence-level tasks.

To obtain a sentence representation from BERT without fine-tuning, it is necessary to pool across tokens and it is recommended to leverage the second-to-last hidden layer[3], which I presently follow, averaging this token-level representation.

5.5 Supervised approaches

5.5.1 InferSent

InferSent [Conneau et al., 2017] is an approach standing out for being trained on two corpora of high quality annotated data. The Stanford Natural Inference corpus (SNLI) dataset [Bowman et al., 2015] comprises 570,000 English sentence pairs. annotated with one of three possible labels: *entailment*, *contradiction* and *neutral*. The authors note that it is one of the largest high-quality annotated resources, which has been curated explicitely with the understanding of sentence semantics in mind. The MultiNLI corpus [Williams et al., 2018] is a multi-genre version of SNLI with 433,000 sentence pairs, spanning ten genres of written and spoken English.

[3] As per a comment by the principal author: `https://github.com/google-research/bert/issues/71#issuecomment-436507081`

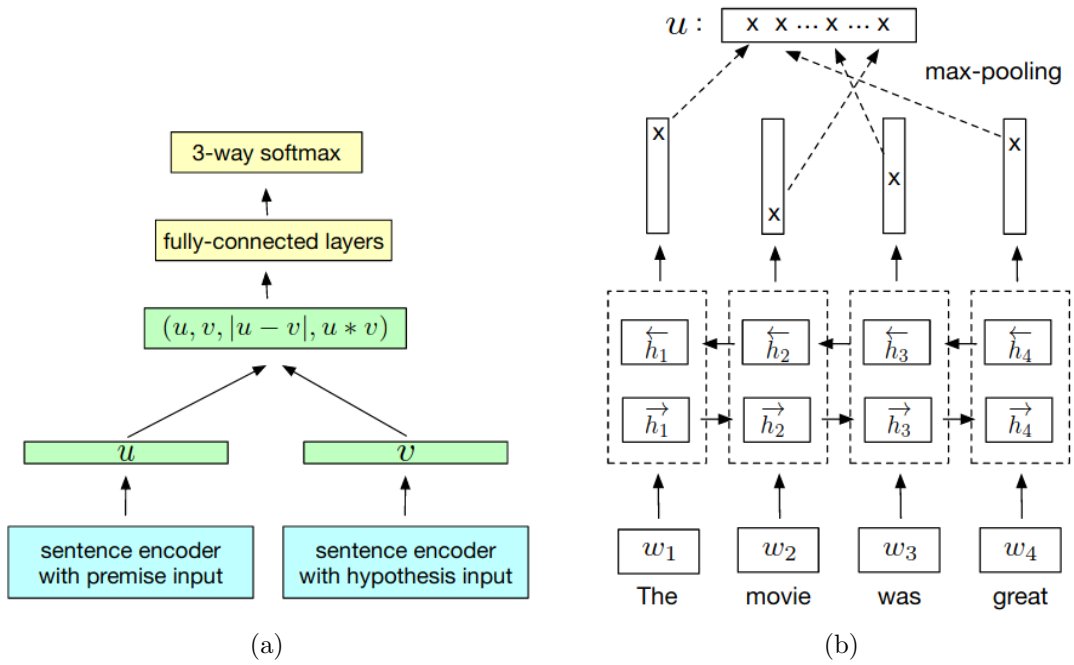


Figure 3: InferSent (a) NLI training scheme (b) Bi-LSTM max-pooling sentence encoder. Source: [Conneau et al., 2017]

The authors contend that as such, it provides coverage for most of the language's complexity.

Figure 3 illustrates the training regime and sentence encoder. For each sentence pair, premise and hypothesis are separately encoded. Relations are represented as a vector which contains the concatenation, element-wise product and absolute element-wise difference of the two vectors. The resulting vector represents information from both sentences. It is the input to a three-class classifier, which consists of several fully-connected hidden layers and a softmax layer. Conneau et al. [2017] evaluate sentence encoder architectures of various complexity, the best performing being a (relatively simple) BiLSTM, which is the concatenation of the outputs of two LSTMs reading the sentences in the forward and backward directions respectively. The number of hidden units depends on the input sequence length, given the recurrent nature of an LSTM. To obtain a vector with fixed size, the authors apply max-pooling over the dimensions of the hidden units.

In their evaluation, the authors note that at the time of writing, the SkipThought-LN model, trained on large corpora of ordered sentences, is the best-performing sentence encoding method. Their approach succeeds in outperforming it in all instances, whith significantly less data (570,000 compared to 64 million sentence). They attribute this to the high-quality supervision found in the SNLI dataset. They conclude that natural language inference data allows sentence encoders to learn uni-

versally useful representations.

5.5.2 Universal sentence encoder

The universal sentence encoder (USE) presented by Cer et al. [2018] has been developed with the goal of obtaining representations that are specifically useful for transfer learning to other NLP tasks. This is motivated by the fact that deep-learning approaches generally require large data sets and annotating supervised training data is highly costly and thus impossible to curate in the context of most NLP tasks, both in research or industry. The authors draw on findings by [Conneau et al., 2017] (InferSent), showing robust performance in transfer tasks when relying on pre-trained sentence embeddings.

Cer et al. [2018] note that transfer learning relying on sentence embeddings generally outperforms word-level transfer. Two different encoders are described, a more powerful transformer-based variant (see Vaswani et al. [2017] for details) and a computationally less expensive variant based on a deep averaging network (DAN) [Iyyer et al., 2015]. The authors note that the former scales quadratically (and thus dramatically) in both model time and space complexity with increasing sentence length, but generally performs better in downstream tasks. The transformer-based model builds sentence embeddings through the encoding sub-graph that is part of the transformer architecture [Vaswani et al., 2017]. This sub-graph relies on attention to obtain word representations that are context aware and consider order and identity of the remaining words.

The DAN is a minimal variation on a feedforward network. Input embeddings for words and bigrams are averaged together prior to passing through several nonlinear layers and a softmax layer, yielding a sentence embedding [Iyyer et al., 2015]. As such, it does not capture context as the transformer variant and is a BoW method. The authors note that the computational complexity of the DAN encoder is linear with respect to the input sequence length, whilst presenting a strong baseline in text classification tasks, sometimes matching or outperforming the more complex transformer version.

The authors emphasize that the encoding model is kept general in that a single encoding model feeds several downstream tasks, in a multi-task learning setting. The tasks are unsupervised learning from running texts in the vein of Skip-Thought [Kiros et al., 2015], an input-response task which models conversation to include parsed conversational data [Henderson et al., 2017], and several classification tasks trained to introduce supervised data. The input to either encoder is lowercased and

tokenized using the Penn Treebank Tokenizer [Manning et al., 2014]. A fixed length 512-dimensional sentence representation is obtained by summing the representations element-wise at each word position and dividing by the square root of the sentence length to account for sentence length effects. Unsupervised training data is obtained from sources such as Wikipedia, web news, question-answer websites and discussion forums. Finally, as with InferSent, the SNLI corpus is used to augment the encoder with supervised data. Contrary to the former approach, the authors do not mention whether inference labels are used. Given the architecture described above, it appears that only the sentence pairs themselves are leveraged.

Cer et al. [2018] evaluate the models on a variety of transfer tasks, such as question classification and sentiment analysis of movie reviews and customer reviews. To this end, the output of the sentence encoder serves as the input to a task-specific DNN. The authors conclude that transfer learning helps improve the performance of many tasks. This is of particular importance when training data is limited. They concede that models not relying on transfer learning approach ones that do, when the training set size is increased.

As with ELMo and BERT, I am only concerned with the generic representation.

5.6 Summary

Approach	Training	Dimensionality
GloVe 50	un-/self-supervised	50
fastText	un-/self-supervised	300
Power-Mean	-	3600
ELMo (all layers)	un-/self-supervised	3072
BERT (2nd-last layer)	un-/self-supervised	1024
Skip-Thought	un-/self-supervised	4800
InferSent	supervised	4096
USE (DAN)	supervised	512
USE (Transformer)	supervised	512

Table 4: Overview of embedding approaches and corresponding training methods and dimensionalities. Adapted from [Perone et al., 2018]

Table 4 briefly summarizes training methods and dimensionality of evaluated embeddings.

The approaches chosen in the context of this evaluation differ considerably in their

architectural complexity, used training data and dimensionality. In several instances, sophisticated and computationally expensive approaches are either matched or outperformed by (sometimes much) simpler methods, methods relying on less training data or representations with lower dimensionality. In [Perone et al., 2018], *fastText* outperforms prior word-level approaches such as GloVe, particularly when leveraging subwords, while allowing rapid training. *Power-Mean* shows the effect of combining embeddings with complementary linguistic features. At the same time, *InferSent* illustrates the effect of high quality data, performing well with several orders of magnitude less training data than found for previous successful approaches, both on the word- and sentence level. Finally, the performance of the transformer-based variant of *USE* is competitive with InferSent in the evaluation carried out by Perone et al. [2018], despite having a dimensionality of just 512 (compared with InferSent at 4096, only exceeded by the combined Skip-Thought model). This shows that very useful information can be represented in what is a relatively low-dimensional vector for sentence embeddings standards. Finally, *BERT* constitutes the most recent and complex approach presently discussed, obtaining state-of-the-art performance on a wide range of tasks upon publishing [Devlin et al., 2018]. As such, it is of particular interest whether this is reflected in the subsequent evaluation.

6 Evaluation

The evaluation was performed using the command-line tool described in Appendix A. An illustrated summary of the process is shown in Figure **??**. Following, I first describe the selected pre-trained embedding models, regression model, pre-processing and evaluation setup. I then summarize high-level results and consider performance with respect to specific linguistic features, as well as features and subject for eye-tracking and fMRI, respectively. Finally, I consider correlations between datasets and modalities, as well as between modalities and extrinsic results.

6.1 Pre-trained embedding models

For practical reasons, not all available model variations are optimized exhaustively with respect to batch and layer size (see below). Instead, better performing variations are substituted upon obtaining the finally obtained parameters, if this improves upon the initially chosen model[1].

I use **GloVe** embeddings with the lowest available pre-trained dimensionality of 50 in this work, which are subsequently called GloVe 50. The embeddings were pre-trained on the Wikipedia 2014 and Gigaword 5 [Parker et al., 2011] datasets (six billion tokens) and are uncased[2]. Contrary to other word embeddings, the intention is not to compare the performance of GloVe, but to establish a lower bound. A corollary of this choice is that the GloVe 50 embeddings are the least likely to suffer from dimensionality problems incurred by the dataset sizes, which are significantly reduced going from the word to the sentence aggregation level. Indeed, GloVe 50 is the only evaluated embedding type for which the ratio between the number of sentences and the dimensionality is always greater than one across all modalities.

[1] While this is a shortcut, I note that [Perone et al., 2018] observe meaningfully differences using a single fixed layer and batch parametrization, which is shared by all approaches tested in both downstream and linguistic probing tasks. It is reasonable to assume that different models obtained using the same general approach do not materially differ in optimal batch and layer sizes. As such, the present approach constitutes a compromise between a fixed parametrization and an exhaustive parameter search.

[2] `https://nlp.stanford.edu/projects/glove/`

For **fastText**, parameters were optimized using fastText without subwords trained on Wikipedia 2017, the UMBC webbase corpus and the statmt.org news dataset (suffixed with Wiki as follows, 16 billion tokens total). For the final parametrization, fastText Wiki without subwords, as well as a model based on the Common-Crawl corpus (suffixed with CC as follows, 600 billion tokens, both with and without subwords) were considered, as well as more recently published models trained on both, including subwords. Due to technical constraints, subwords could not be leveraged for the separately trained models [3]. Only for the most recent and largest model trained on both corpora, subwords are considered (suffixed with combined as follows). For EEG, fastText Wiki with subwords was found to perform best, despite subwords not being used. For eye-tracking fastText Wiki with subwords is used for Dundee, fastText CC without subwords for GECO and fastText CC with subwords for ZuCo. For fMRI, fastText CC without subwords was selected for the smaller Pereira dataset and fastText Wiki with subwords for the remaining two. Unexpectedly, fastText combined, the only model actually leveraging subwords during embedding lookup, was not selected for any dataset. This suggests that in predicting cognitive signals, the significantly greater amount of training data available for the models trained on Common Crawl data is not universally beneficial and subwords do not have a significant impact for the chosen corpora.

For **Power-Mean**, I use monolingual TensorHub model provided by the authors[4].

For **ELMo**, parameters are optimized for the default model trained on the *1 Billion Word Benchmark*[5]. For the final parametrization, the larger pre-trained model trained on 5.5 billion tokens[6] performs better for most datasets (in line with results observed by Perone et al. [2018]) and is substituted accordingly[7]. An exception is the GECO dataset, where a slightly better average error is obtained for the default model.

For **BERT**, the pre-trained models used in this evaluation rely on the BooksCorpus [Zhu et al., 2015] (see above for statistics) and an English Wikipedia corpus (2,500M words)[8]. Only the most recent variant of the large model with improved (whole) word masking is evaluated, given that it outperforms the previous model[9]. A sentence representation is obtained by mean-pooling the representations of the

[3] `https://fasttext.cc/docs/en/english-vectors.html`

[4] `https://github.com/UKPLab/arxiv2018-xling-sentence-embeddings`

[5] `https://opensource.google/projects/lm-benchmark`

[6] Wikipedia (1.9B tokens) and monolingual news crawl data of WMT 2008-2012 (3.6B tokens)

[7] `https://allennlp.org/elmo`

[8] https://github.com/google-research/bert

[9] `https://github.com/google-research/bert`

second to last layer for a maximum sequence length of 100. For eye-tracking, the uncased model is selected and for EEG the cased model. For fMRI, Pereira (small) and Wehbe, I select the uncased model and the cased model for Pereira (large).

For **Skip-Thought**, I do not rely on the original Skip-Thought implementation for technical reasons, but draw on an inofficial re-implementation in TensorFlow[10] instead. I use a pre-trained model based on the same data used by the original authors (BookCorpus with approximately 985 million tokens and 74 million sentences, from 11,038 books [Zhu et al., 2015]).

For **InferSent**, two pre-trained variants are available, trained with GloVe and fastText vectors respectively[11]. The fastText-based model outperforms the GloVe-based model in most *SentEval* tasks and is chosen accordingly for the present evaluation. I note that the authors chose fastText CC without subwords as their basis for this model.

For **USE**, parameters are optimized using the DAN variant[12]. With the final parmetrization, it performs identically or better than the Transformer-variant[13] for most datasets, except for Pereira (large) and Wehbe, where the latter performs minimally better and I choose it accordingly. This contrasts with previously discussed extrinsic results, which clearly favor the Transformer-based approach in most instances.

6.2 Regression model

A simple feedforward network (multi-layer perceptron, or MLP, Rosenblatt [1962]) is used as basis for the model architecture. The model specification consists of an input layer matching the dimensionality of the respective embeddings, two identically sized hidden layers in all experiments and an output layer where the dimensionality corresponds to the respective cognitive dataset. As a regressor, it predict continuous values, contrasting with discrete labels predicted by a classifier. I use a fixed dropout of 0.5 after each hidden layer (which has been found to maximize regularization by Baldi and Sadowski [2013]) and batch normalization after the last. All layers use the *relu* activation function, except for the output layer, which is linear. The model

[10] Community implementation of Skip-Thought vectors in tensorflow by Chris Shallue: `https://github.com/tensorflow/models/tree/master/research/skip_thoughts`

[11] `https://github.com/facebookresearch/InferSent`

[12] `https://tfhub.dev/google/universal-sentence-encoder/2` (Tensorflow 1.x model)

[13] `https://tfhub.dev/google/universal-sentence-encoder-large/3` (Tensorflow 1.x model)

is implemented in Keras API[14] as a regressor. As loss function I use the *mean squared error* (MSE) and train the model using the Adam optimizer [Kingma and Ba, 2014], which has been shown to perform robustly across many tasks. Keras default parameters are used for Adam (learning rate 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-7}$).

6.3 Pre-processing

6.3.1 Sentence pre-processing

For GloVe 50 and fastText, I consider (proper) nouns, adjectives that are not (separate) affixes, verbs that are not modal and adverbs that are not particles or "wh"-adverbs (*when*, *why*, etc.). The cased form of a word is preferred if present, otherwise a fallback to the lowercased occurs. If neither can be found, the word is assigned the zero vector. For ELMo, the entire sequence is considered during averaging in order to retain the context information. For Power-Mean, the used model takes a whitespace-tokenized sentence as input, therefore I do not filter tokens here. The authors note that input has to be lowercased when relying on this model. However, I obtain lower average errors for all datasets but Pereira (small) when retaining casing and thus only lowercase for the latter dataset. For averaging baselines and Power-Mean, tokenization and POS tagging (where applicable) is performed using `spaCy`[15], relying on the largest English model[16]. All other approaches rely on built-in tokenization, and casing is retained.

6.3.2 Feature selection and transformation

When performing scaling operations on an entire dataset, data leakage may occur, leading to unwanted inference of unseen data by the model. To avoid this issue with respect to the dependent or target variable, any pre-processing step influencing the scaling of cognitive vectors is performed within the inner and outer cross-validation loop. In the given setting, this is mainly to prevent inferred information from overpowering the predictive power of individual embeddings, artificially lowering errors and hindering comparison between tested hypotheses.

[14] `https://keras.io/`

[15] `https://spacy.io/`

[16] `https://spacy.io/models/en#en_core_web_lg`

An exception to this on-the-fly transformation is made for the pre-selection of voxels for the fMRI modality. This selection has been found to be a technical requirement for performing Kernel PCA on standard-scaled fMRI data (but is performed for all fMRI datsets, for simplicity). The reason is that neighbouring fMRI voxels are highly correlated with one another, which can cause the Kernel PCA to fail due to eigenvalues of the input matrix being negative. To avoid this issue, a fixed number of 15,000 voxels is first randomly sampled to reduce space and time complexity of correlation analysis (using the random seed specified in the configuration for reproducibility). Following, for pairs of voxels with a Pearson's correlation of more than 0.95, one of the voxels is discarded, removing several thousand features depending on the individual dataset. I perform the selection on the entire dataset, with the rationale that learned models are not required to generalize to additional data in the given setting [17]. Furthermore, the selection does not introduce data leakage, as it leaves values and ranges of remaining features unchanged.Finally, it leads to a greater improvement in quality, which is in line with the purpose of the datasets as quality gauges.

For fMRI and EEG data, dimensionality reduction is necessary as for both datasets, the number of input dimensions greatly exceeds the size of datasets. As it has been found to improve scores, features are standardized first, with the exception of the Wehbe dataset, where doing so leads to errors [18]. For fMRI, Gauthier and Levy [2019] have shown that PCA can capture most of the variance in as little as 256 dimensions, which I adopt here. PCA performs singular value decomposition to project data to a lower dimensional space [Wold et al., 1987]. I apply Kernel principal component analysis (Kernel PCA), which is a extension of linear PCA [Schölkopf et al., 1997]. The usage of a kernel function allows projecting data into a higher dimensional space, where non-linearities become linearly separable (kernel trick). Kernel PCA performs better than linear PCA for both modalities, indicating that non-linear features are present. The final parametrization specifies for both modalities at most 256 dimensions[19], a polynomial kernel, a gamma value of 0.01 and a degree of 3 (independently determined for either modality). I note that for practical reasons, no extensive hyperparameter search was carried out to optimize the representation of the cognitive data.

[17] Which in the case of fMRI is not possible when relying on subject-level datasets, which constitute separate hypotheses due to differences in brain morphology.

[18] I note that this indicates underlying issues with this dataset, potentially stemming from the averaging procedure performed to obtain sentence representations

[19] The dimensionality is dynamically clipped to the highest possible value depending on the training fold size of the corresponding CV loop, as it needs to be strictly less than the number of data points. All fMRI datasets as well as the ZuCo 2 EEG dataset require clipping of the PCA dimensionality due to their size, in either the inner or both CV loops.

Finally, the resulting matrix is Min-Max-scaled between 0 and 1 on a per-feature basis, to allow interpretation of MSE scores without specific knowledge of feature ranges. For eye-tracking, only Min-Max-saling is performed, as the dimensionality is sufficiently low in all instances (see Section 4.1). As the scaler is fitted only on the training portion of the targets, values falling outside its range may occur in the test set, leading to negative values or values greater than 1. These values are clipped at 0 and 1 respectively, to ensure that the obtained MSE remains scaled between 0 and 1. This leads to an information loss, however I argue that it is compensated by the cross-validation procedure[20].

6.4 Experimental setup

Only the number of hidden layers, layer sizes as well as batch size were optimized. Activation (relu), number of folds for the inner and outer cross-validation (CV) loop (3 and 5, respectively) and number of epochs (50) was fixed in all cases. Two hidden layers were found to outperform a single hidden layer in initial tests. When optimizing hidden layer sizes, I started with a parameter selection spanning 25% to 75% of the input embedding dimensionality and subsequently expanded it in instances where a boundary value was selected. ,

A nested cross-validation (CV) setup is used, with the hyper-parameter search taking place in the inner *three*-fold loop and the actual prediction occurring in the outer *five*-fold loop in order to obtain a robust estimation of model fit. The model evaluated in the outer loop is trained by refitting the model on the entire dataset used for the inner loop, using the best parameters obtained in it. No separate hold-out (test) set is reserved, as the small dataset sizes already lead to very sparse training fold sizes (particularly in the inner CV loop) and the model is not expected to generalize to additional data given the artificial nature of the task.

For the final parametrization, the layer size selection most frequently picked throughout the outer CV is selected, as soon as the parameter converges. I assume convergence as soon as a non-boundary value is chosen of the respective grid or an extremum is picked, i.e. either single-digit layer sizes (observed for fMRI) or a layer size corresponding to 90% of the input dimensionality. Batch sizes were evaluated less exhaustively (and prior to layer optimization), with values being multiples of 8 and the best value chosen after approximately two to four searches. For fMRI, I use the same parametrization for all subjects. Search is confined to the first three

[20] Throughout folds, different subranges of the range of a feature are clipped, the union of which necessarily corresponds to the full range of the feature.

subjects per dataset, to reduce computational cost, yet still obtain a "majority vote" in most instances. If results differ on all three subject, the middle value is selected. Train-validation loss plots are examined to assess training history. In all cases, both losses converge with a significant margin before the final epoch and no overfitting is observed. The validation set was eliminated from final run to increase the training data available in the inner CV.

For eye-tracking and fMRI, results are averaged among hypotheses, i.e. features (eye-tracking) and subjects (fMRI), respectively. Results are further averaged to obtain averaged performances as well as significance ratios per modality. To allow significance testing, a set of ten random baselines (randomly generated vectors), is generated per embedding type, using different random seeds and matching the dimensionality of the embeddings. To obtain a robust estimate of random baseline performance, the average of the errors of the individual baselines is computed, as obtained when predicting a cognitive dataset with the same hyperparameters as the embeddings.

All computations were performed on CPU, which was possible given that individual experiments ran in a relatively short time even when using only one core. Given that each worker can take up multiple GBs of RAM, this also allowed running ten or more workers in parallel.

6.5 Results

Given that all reported metrics are measures of error, lower values are better in all instances. Table 5 shows aggregated mean squared error values (MSE) per modality showing that absolute differences are minimal except for eye-tracking. However, a clear ranking is apparent, which I discuss as follows. The MSE is a standard metric of estimator quality for continuous variables, averaging the square of the differences between predicted and reference values. In the following subsections, I discuss the results on the level of modality and dataset. Tables 10 to 12 in Appendix C show averaged MSE per dataset.

Final hyperparameters are shown in Tables 29 through 33 in Appendix C. Limited by the scale of the evaluation, the parameter search can by no means be considered exhaustive, however I argue that the magnitude of error values and potential correlation between datasets and modalities are good indicators of the validity of the results.

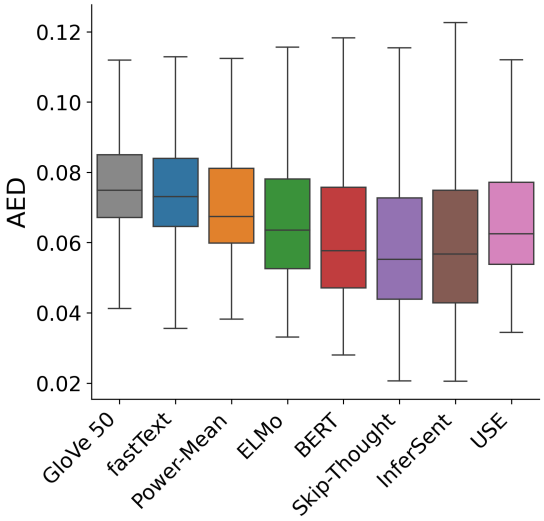
For EEG, selected hidden layer sizes range between 2.5% and 50% of the input size.

	Eye-Tracking (13)		EEG (3)		fMRI (21)	
	Baseline	**Embed.**	**Baseline**	**Embed.**	**Baseline**	**Embed.**
GloVe (50)	0.0181	0.0171	0.000405	0.000103	**0.00426**	0.00430
fastText	**0.0172**	0.0165	0.000812	0.000097	0.00439	**0.00412**
Power-Mean	0.0200	0.0143	0.000259	0.000102	0.00429	0.00416
ELMo	0.0203	0.0130	0.000164	0.000105	0.00438	0.00423
BERT	0.0204	**0.0123**	**0.000133**	0.000110	0.00437	0.00448
Skip-Thought	0.0202	**0.0123**	0.000405	0.000096	0.00462	**0.00412**
InferSent	0.0202	**0.0123**	0.000554	0.000094	0.00464	**0.00412**
USE	0.0195	0.0147	0.000233	**0.000093**	0.00483	**0.00412**

Table 5: Aggregated average MSE on the modality level for embeddings and corresponding random baselines. Results are averaged across the five CV folds of a hypothesis, as well as across hypotheses (features and subjects for eye-tracking and fMRI, respectively). Lowest (best) errors are highlighted per column. The value in parentheses specifies the number of hypotheses per modality.

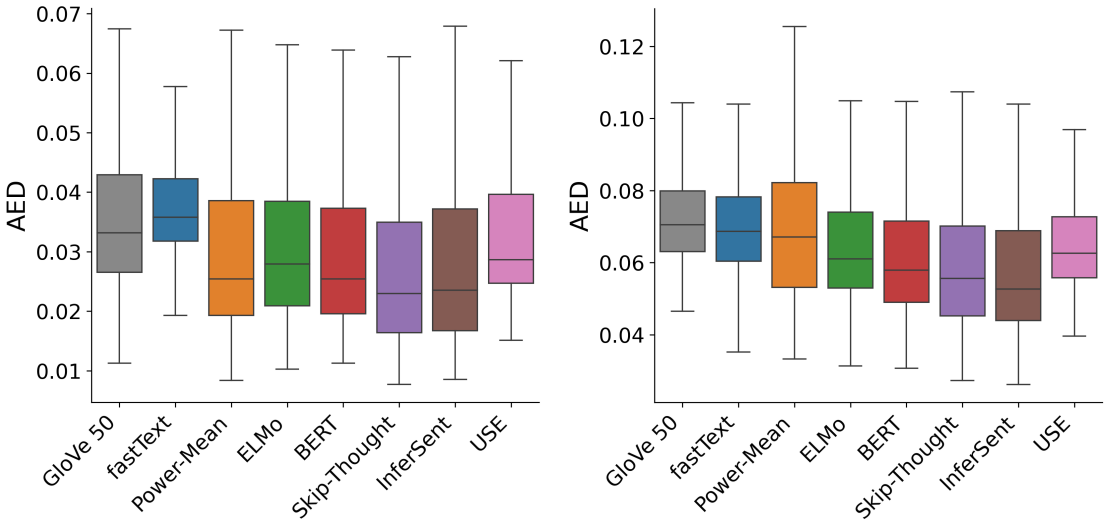
Counterintuitively, the best-performing hidden dimensionality is significantly lower for fMRI, with many complex approaches having a final layer size ranging between five to seven (no lower values were evaluated). This indicates that at least at the given dataset size, relatively fine-grained features can be obtained from the indirect EEG signals, while for the originally very complex and high-dimensional fMRI data, only coarse high-level features are extracted.

To get an understanding how error values are distributed, results are subsequently also presented as box plots. The plots show the median, the inter-quartile range (IQR, represented by the box), corresponding to the distribution of the middle half of observed errors, or the difference between the 75th and 25th percentile, as well as minimum and maximum (excluding individual outliers). The fact that MSE heavily weighs outliers (due to the squaring) make interpretation somewhat more difficult. For this reason, the plots show the distribution of the **absolute error averaged across dimensions** over sentences for each dataset (which I informally refer to as AED). This is a non-standard metric accounting for the high dimensionality of the target variable. I note that neither MSE nor AED account for sentence length. As such, very short sentences may only contain only one or even no distinct semantic propositions, yet contribute to the metric to the same extent as a sentence with richer semantics.



(a)

Figure 4: Distribution of sentence-level AED for **Dundee**.



(a) (b)

Figure 5: Distribution of sentence-level AEDs for **GECO** (a) and **ZuCo** (eye-tracking (b).

6.5.1 Eye-tracking

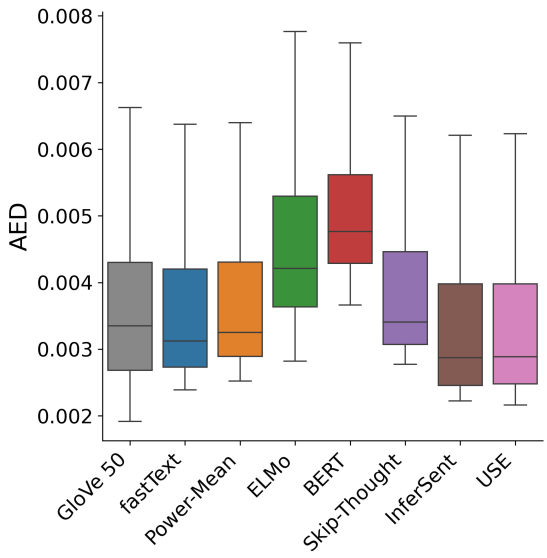
Figures 4 and 5 show AED box plots for individual datasets and Table 10 in Appendix C gives an overview of average MSE values of random baselines and embeddings per dataset.

Across datasets, the best-performing random baseline (300 dimension) obtains an average MSE of 0.0172, while the best-perfoming embeddings yield 0.0123, compared with 0.0202 - 0.0204 for the random baselines of the same dimensionality.

As noted in Chapter 2, eye-tracking is the most indirect method of measuring cognitive processes and offers the least rich representation, with a single scalar obtained per word and feature. As such, the clear differences in ranking on the dataset level and partially on the modality level, are somewhat unexpected. The most plausible reasons for this are, on one hand, the size of datasets, which in the case of ZuCo are at least identical to the EEG datasets, and significantly exceed all other datasets in the case of Dundee and GECO. On the other hand, the data has a significantly lower dimensionality which is bounded by the maximum sentence length and makes lossy dimensionality reduction unnecessary. This also means that eye-tracking is the only modality for which sequence information is explicitely retained, whilst this is unclear for EEG given the application of Kernel PCA to the concatenated word-level vectors and definitely not the case for fMRI, where I use averaged sentence-level representations.

Aggregated across all three datasets, GloVe 50 shows the highest mean error, which is expected. BERT, InferSent and Skip-Thought are tied for the lowest error, with ELMo falling minimally behind. I also note the small error range, making interpretation difficult. With the exception of GloVe, approaches are clearly offset from their corresponding random baselines. When considering AED scores, Skip-Thought and InferSent are more or less tied for the lead for all three datasets, showing a slightly lower median error than BERT. ELMo and USE fall behind to varied degree, approaching random baselines.

While the differences in errors are generally small, this indicates that approaches relying on recurrent sentence encoders yield representations which are more predictive of cognitive signals, outperforming context-sensitive averaging baselines such as ELMo, as well as USE, which performs competitively in various downstream tasks. Unexpectedly, the most recent and competitive approach BERT does not lead the ranking.



(a)

Figure 6: (Distribution of sentence-level AED for **Natural Speech**.

6.5.2 EEG

Figures 6 and 7 show AED box plots for individual datasets and Table 11 in Appendix C gives an overview of average MSE values of random baselines and embeddings per dataset. Across datasets, the best-performing random baseline (1024 dimension) obtains an MSE of 0.000133, while the best-perfoming embedding, USE, yield 0.000093, compared with 0.000233 for the random baseline of the same dimensionality (512).

The very low average MSE values indicate that despite the small size of datasets and higher dimensionality of the target data, the regression model succeeds in learning the problem with small error regardless of the input embedding. All approaches outperform random baselines, however absolute differences between embeddings are very small.

Notably, the worst performing random baseline still achieves an average MSE of 0.000812, aggregated across all three datasets and is lower than the baseline errors observed by Hollenstein et al. [2019] on the word-level. This shows that scales cannot be compared between datasets. Ont the dataset level, random baseline error ranges vary similarly to the embeddings (although with greater magnitude), with ZuCo showing the lowest errors, followed by Natural Speech and ZuCo 2, for which the difference between random baselines and embeddings is most pronounced (see Appendix C).

Inspecting the spread of the mean absolute error (AED) yields clear differences between approaches and obtains results that differ markedly from eye-tracking. For

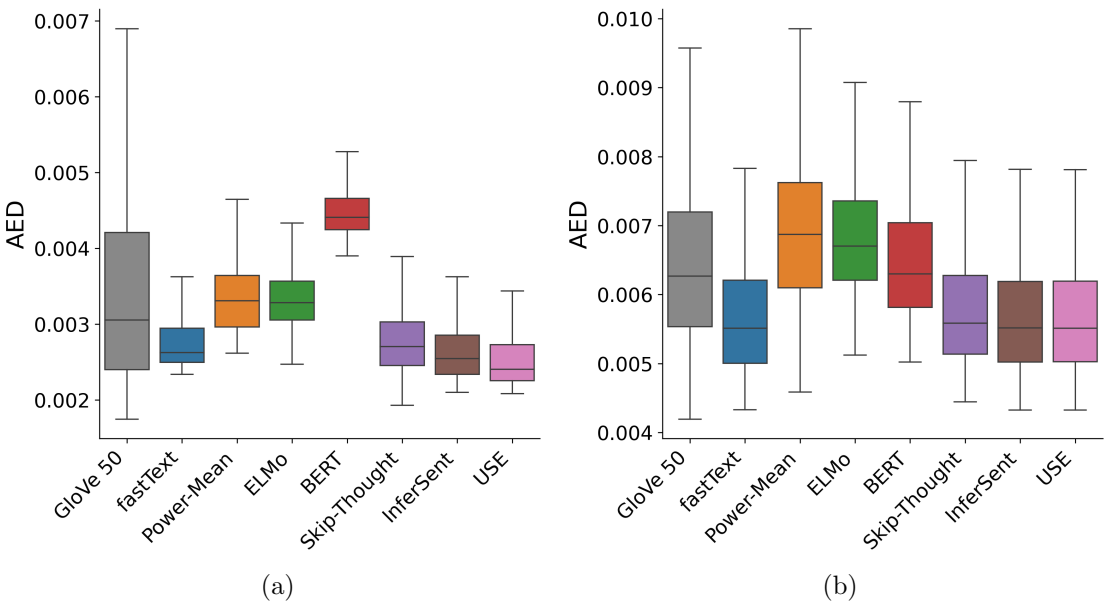
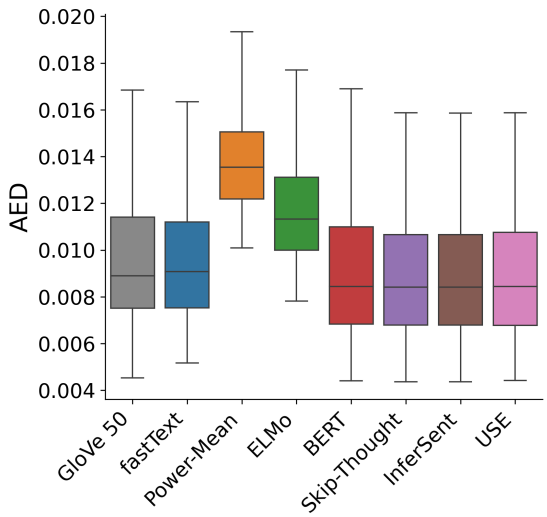


Figure 7: Distribution of sentence-level AEDs for **ZuCo** (EEG) (a) and **ZuCo 2** (EEG) (b).

all datasets, ELMo falls clearly behind the worst-case baseline GloVe 50 with respect to the median error, as does Power-Mean for the ZuCo corpora. USE, on the other hand, obtains the lowest errors on all three datasets, tied with InferSent on Natural Speech and ZuCo 2.

Notably, the three baseline embeddings are only minimally falling behind dedicated sentence encoders and in some cases match them. Conversely, the contextualized embeddings ELMo and BERT fall behind clearly, only outperforming a baseline approach in one instance and thus presenting a strong anomaly. Skip-Thought, which outperformed USE on eye-tracking and generally matched InferSent, now falls moderately behind on two datasets. Finally, fastText ranks unexpectedly well, clearly outperforming GloVe 50 for both ZuCo datasets, minimally outperforming Skip-Thought for Natural Speech and ZuCo, and matching InferSent and USE for ZuCo 2.

The differences in ranking clearly indicate that eye-tracking and EEG leverage different features of input embeddings. The presence of an intervening variable such as the mode of recording or the text material, is made unlikely by three aspects: All three EEG datasets exhibit a roughly similar ranking pattern (although skewed in different ways), despite the fact that Natural Speech relies on narrative text and the ZuCo corpora contain isolated sentences. Furthermore, for Natural Speech, brain activity was recorded while listening to spoken language, whereas for the ZuCo corpora, the text was read in a self-paced manner. Finally and most notably, the textual part of the ZuCo dataset is shared between eye-tracking and EEG, yet a



(a)

Figure 8: Distribution of sentence-level AED for **Pereira (small)**).

very different ranking is obtained for either modality. Furthermore, the clear difference between embeddings and random baselines strongly suggests that the observed pattern is not arbitrary.

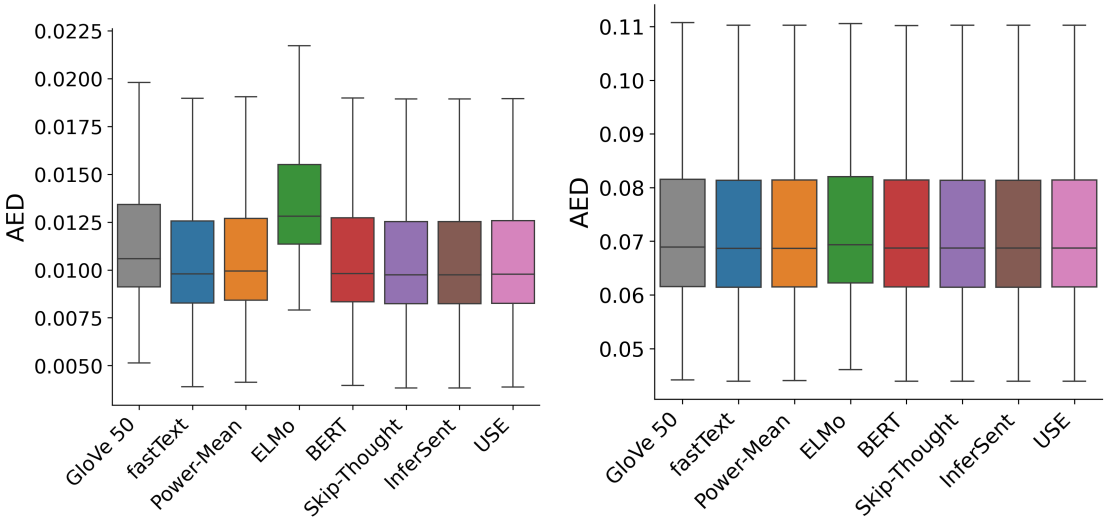
It is possible that a different approach to obtaining a sentence representation from ELMo and BERT may result in a lower prediction error. However, given the middle ranking of those approaches for eye-tracking, it is unlikely that the placement for EEG can solely be attributed to issues relating to the representation.

I concede that gauging differences between evaluated hypotheses is particularly challenging for EEG, given the small scale of errors. Compared with eye-tracking, the small dataset sizes lead to more pronounced range clipping when performing target transformation, which can only be mitigated by curating EEG datasets of similar size.

6.5.3 fMRI

Figures 8 and 9 show AED box plots for individual datasets and Table 12 in Appendix C gives an overview of average MSE values of random baselines and embeddings per dataset. Across datasets, the best-performing random baseline (50 dimensions) obtains an average MSE of 0.00426, while the best-perfoming embeddings, Skip-Thought (tied with InferSent), yield 0.00412 compared with 0.00462 - 0.00483 for the random baselines of the same dimensionality.

As previously noted, fMRI is the most direct access to brain activity and offers the highest spatial resolution. On the other hand, the dataset sizes of this modality



(a) (b)

Figure 9: (a) Distribution of sentence-level AEDs for **Pereira (large)** and (b) **Wehbe**

are by far the smallest, which is likely an important factor for the results observed here. Averaged across all three datasets, differences in averaged MSE are minimal between datasets and whilst random baseline medians are outperformed, the margin is generally very small. For Pereira (small), Power-Mean and ELMo show the highest error, whilst Skip-Thought, InferSent and USE are nearly exactly tied for the lowest error value. GloVe 50 and fastText place slightly behind, with similar median errors, agreeing with results observed for eye-tracking.

For Pereira (large), even fewer differences are apparent, with ELMo again placing last, followed by GloVe 50 and near identical erors for all remaining approaches. Finally, no usable signal is obtained from the Wehbe dataset. Given that only the smallest dataset shows a clear ranking, the evidence for this modality is not conclusive. However, the ranking for Pereira (small) is plausible, with more sophisticated approaches outperforming averaging baselines such as GloVe 50 and fastText, suggesting, that the results are generally valid.

More generally, results suggest the presence of a lower bound for the error, which is likely related to the small dataset sizes. As previously noted, final layer sizes selected for fMRI during hyper-parameter search are unexpectedly small, indicating that only a few coarse, high-level features can be learned. Counterintuitively, I observe lower errors and a more defined ranking for Pereira (small). This is despite nearly identical pre-processing, with only the PCA dimensionality differing due to the difference in dataset size. When matching the PCA dimensionality of Pereira (small) (193), scores deteriorate further. It is possible that this is related to the

narrative portion of the passages, given that the remaining text material is similar to Pereia (large) and subjects overlap between the two experiments. This may also contribute to the difference in Power-Mean performance. On the other hand, narrative passages only constitutes a third of a total of 243 sentences. As such, it is also possible that Pereira (large) has a more dominant noise component that the present evaluation setting cannot compensate. For Wehbe, Hollenstein et al. [2019] observed small but relatively clear differences on the word level. It therefore appears that simple sentence-level averaging of scans deteriorates signal quality beyond usefulness. I do not exclude the possibility to obtain a useful respresentation using more sophisticated means of aggregating sentence-level information.

An important aspect of raw fMRI data is its very high dimensionality. However, for Pereira (small), even after randomly sampling 15,000 voxels (roughly a tenth of observed voxel counts), approximately 10% voxels on average a show correlation of more than 95% (and are discarded), which is to be expected given that fMRI has a much more fine-grained spatial resolution than EEG and neighboring voxels necessarily have a similar level of activation. This indicates very high information redundancy, as supported by Hollenstein et al. [2019]'s findings that as little as 1000 randomly sampled voxels suffice to differentiate embeddings on the word level. At the same time, the very high explained variance I observed when initially evaluating linear PCA on the entire datasets with a fixed dimensionality of 256 (at least 94%), as well the additional improvements observed when shifting to a kernelized approach do suggest that the obtained vectors are a rich representation of the input data.

6.5.4 Overall embedding performance

The results described above yield a very heterogenous image on the embedding level.

As expected **GloVe 50** and **fastText** perform at the bottom end for eye-tracking. However, fastText only gains minimally over GloVe 50, which is surprising given that GloVe 50 is artificially penalized due to its much lower dimensionality and I selected the best-performing fastText model among five. Contrasting with this, fastText outperforms GloVe 50 on all EEG datasets, also yielding a smaller error spread for two datasets. For this modality, fastText is also highly competitive with sentence encoders, slightly outperforming Skip-Thought for all datasets. For fMRI, fastText outperforms GloVe 50 only on Pereira (large). This indicates that additional information encoded by fastText vectors only affords a marked improvement when predicting EEG data. **Power-Mean** performs similarly to fastText for two eye-tracking datasets, whilst clearly outperforming it for GECO, which may be related to the fact

that this is a single-document dataset (novel). For EEG, Power-Mean clearly falls behind fastText for two datasets, and even behind GloVe 50 for ZuCo2. For fMRI, it clearly obtains the largest error on Pereira (small) and minimally underperforms fastText on Pereira (large). This is again quite surprising, given that Power-Mean is quite competitive with InferSent in extrinsic evaluation. The contextualized **ELMo** and **BERT** embeddings differ particularly strongly between modalities, whilst often obtaining rankings similar to each other. For eye-tracking, they generally assume a middle rank, outperforming the previously discussed baselines (except for GECO), but unexpectedly falling behind older sentence encoders. The benefit of BERT over ELMo is in line with extrinsic findings such as presented in [Devlin et al., 2018]. Surprisingly, the approaches assume low ranks for EEG. For fMRI, only ELMo falls behind, with BERT matching other approaches. The eye-tracking performance indicates that the contextual information encoded by the approaches affords a modest advantage when predicting this modality. The recurrent **Skip-Thought** and **InferSent** sentence encoders achieve the lowest errors across modalities, with the former lagging slightly behind and only falling back for the Natural speech dataset. This underlines the usefulness of sequence-aware sentence encoding using recurrent neural networks and indicates that these approaches yield richer representations that are predictive of more cognitive features than any other tested approach. Finally, **USE** (DAN) performs inconspicously for eye-tracking, slightly underperforming ELMo. For EEG however, the approach is very competitive with the recurrent encoders, outperforming Skip-Thought and either matching or exceeding InferSent in all instances. This continues through fMRI with a near identical median for both Pereira datasets. Given that USE DAN is a BoW approach, its eye-tracking performance strongly indicates that this modality benefits from the sequence information encoded by Skip-Thought and InferSent. Conversely, USE dispels the notion that high-dimensional representations are necessary in all instances, matching or outperforming the recurrent encoders for the other two modalities, despite having only a dimensionality of 512. The similar performances of InferSent and USE (DAN) for EEG also indicates that the usage of high quality inference data benefits disparate approaches, especially as the otherwise competitive Skip-Thought falls behind on two of three EEG datasets.

In summary, the findings strongly suggest that that modalities measure different properties of embeddings, although for fMRI, the picture is presently incomplete. Recurrent approaches to sentence encoding appear to capture a wide range of measured properties, while some BoW approaches yield comparable errors for EEG data, but fall behind clearly with respect to eye-tracking. Notably, the role of Power-Mean and the contextualized approaches remains unclear, as the observed

performance partially contradicts extrinsic results, particularly for EEG and partially for Pereira (small). Finally, the results observed for eye-tracking suggests that sequential features can be learned from the input representations, which is itself an intact sequence as it is not subject to dimensionality reduction as the other modalities. However, given that the regressor is not a sequence-to-sequence model (see [Sutskever et al., 2014]), such information can only be inferred from combinations of individual features.

6.5.5 Fine-grained analyses

6.5.5.1 Linguistic features

Following, I consider the relationship between performance and specific linguistic features. This is done by selecting sentences according to the the occurrence of various grammatical dependencies, named entities, combined appearance of specific POS tags, as well as separating sentences into three sentiment classes[20]. As none of these subsets yielded clear differences with respect to the best-performing approach of any dataset, I omit results at this point. For mean AED values and frequencies of occurrences per selected feature, see Appendix C, Tables 14 to 22 .

The absence of differences indicates that observed errors cannot be clearly attributed to the occurrence of specific linguistic features or the polarity of sentences. As such, I contend that observed differences in ranking are attributable to general improvements afforded by the better-performing embeddings, which can not easily be divided into individual factors. This is intuitive, given that a feature's contribution to the error is confounded by the presence of all other features in a sentence. I thus omit a detailed discussion of these findings.

6.5.5.2 Feature- and subject-level analysis

For multi-hypothesis datasets, the above graphs show the aggregate AED averaged across features and subjects respectively. To ensure that this does not obfuscate differences between individual hypotheses, I also consider individual hypotheses. As again, no clear discrepancies are apparent, I show results in in Appendix C, Tables 23 through 25 for eye-tracking and Tables 26 to 28 for fMRI.

[20]For this purpose, the `TextBlob` package (`https://textblob.readthedocs.io/`) was used, which yields a polarity value between -1 and 1 for a sentence. As a simple heuristic, polarity values between -0.25 to 0.25 were mapped to neutral (to allow for some tolerance), whilst values below and above were mapped to negative and positive, respectively.

For Pereira (small), subjects obtain very similar error values for the best-performing approaches. Most subject show little variation with the exception of GloVe 50. A single yields error values with an overall lower scale. For Pereira (small), variations are more pronounced, however in many cases, clear preference is given to the approaches obtaining the lowest error on the averaged level. For Wehbe, the AED is very similar across embeddings for all subjects. Between eye-tracking features, clear differences in scale are apparent, however not between embeddings tested against any specific feature.

In summary, an analysis of individual eye-tracking features and fMRI subjects indicates that some variation is present and differences in scaling are observable between features. This is in line with findings by Hollenstein et al. [2019], who note that more general eye-tracking features such as total reading time and the number of fixations encompassing the entire reading process appear to be easier to predict. However, the averaged results neither obfuscate meaningful subject-level differences, nor differences in how well individual approaches predict specific eye-tracking features.

6.5.5.3 Spread of MSE for different models of an approach

As a byproduct of selecting the best-performing fastText model among five, it is possible to assess the range of MSE values between models. Figure 10 shows that for this approach, differences in the mode of training (with or without subwords) and/or the underlying training data lead to a greater spread of propotionally scaled error for EEG than the other modalities. This is further evidence that EEG leverages different features from embeddings and is more sensitive to the mode of training and/or training data than other modalities. However at this point, it is not possible to examine this in further detail.

6.6 Statistical significance testing

In order to determine whether the performance of embeddings is significantly different from that of baselines, statistical hypothesis testing is necessary.Hollenstein et al. [2019] account for two aspects, which also apply to the present evaluation. The first is that the likelihood of obtaining spurious results increases with the number of inferences considered simultaneously (multiple comparison problem). The second is that the distribution of the test data is unknown.

To account for the first issue, the authors consider the global null hypothesis, which

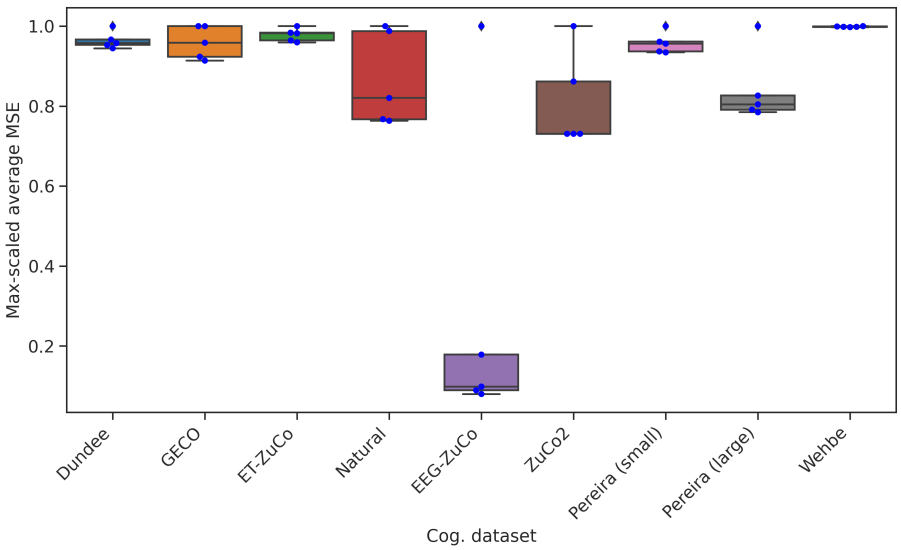


Figure 10: Distribution of max-scaled MSE scores obtained using the final parametrization of fastText for the model variants Wiki (with /without subwords), CC (with/without subwords) and Combined (with subwords). Boxplots are added as visual aids, actual data points in blue.

when rejected, shows that at least one alternative hypothesis is true [Dror et al., 2017]. Following [Hollenstein et al., 2019], I apply the conservative Bonferroni correction, which corrects α (the significance level) such that this null hypothesis is rejected if $p < \frac{\alpha}{N}$, with N being the number of hypotheses[Dror et al., 2017]. In the present evaluation setting, $\alpha = 0.01$ and $N = 3$ for EEG, corresponding to the three datasets, $N = 21$ for fMRI as the total number of subjects across all fMRI datasets [21], and $N = 13$ for eye-tracking, for the total number of chosen features across all eye-tracking datasets.

To account for the second issue, [Hollenstein et al., 2019] perform a two-sided Wilcoxon signed-rank test. The test compares two matched samples (embeddings and baselines in this case) and as its null hypothesis, differences are assumed to follow an asymmetric distribution around zero. The test ranks absolute values of the differences and propagates the sign of each difference to its rank. The test statistic is then the sum of the signed ranks [Dror et al., 2018]. The Wilcoxon signed-rank test only considers the rank of values and is, more generally, a sampling-free test, which scales well on larger datasets. [Dror et al., 2018] note that the statistical power of sampling-based tests is higher, as evaluation measures are directly considered. Such tests compensate for the lack of distributional assumptions with re-sampling, which is more computationally expensive but feasible for small datasets as in the present

[21] I reiterate that the subjects of Pereira (small) are a subset of Pereira (large). However, as the two datasets constitute separate experimental runs, I consider the participation of a single real-world subject in two distinct experiments to yield two hypotheses, to obtain the most conservative estimate.

evaluation.

Nearly all hypotheses test significant under Wilcoxon, therefore I also perform a one-sided paired permutation test[22]. It estimates the distribution of an arbitrary test statistic (here the mean) by computing ideally all possible ways of swapping pairs between the matched samples. In practice, the number of permutations is fixed, approximating the real distribution. For a one-sided test where the test statistic of the first sample is expected to be higher, the difference of the test statistic between the first and second sample is computed first. Applied to the present case, this is the difference between the mean of the inverted error of embeddings and baselines, as the former is expected to be higher (better) than the latter. The p-value is the proportion of permutations where this difference is more extreme than the initially obtained value.

	Eye-Tracking	**EEG**	**fMRI**
GloVe 50	7/13	3/3	20/21
fastText	12/13	3/3	21/21
Power-Mean	13/13	3/3	20/21
BERT	13/13	3/3	20/21
ELMo	13/13	3/3	21/21
Skip-Thought	13/13	3/3	21/21
InferSent	13/13	3/3	21/21
USE	12/13	3/3	21/21

Table 6: Significance ratios of hypotheses per modality, as obtained through the one-sided paired permutation test. $\alpha = 0.01$ with Bonferroni correction per modality. Best ratios are bold, second-best underlined.

The ratios of significant hypotheses per hypothesis is shown in Tables 6 (permutation), as well as 13 (Wilcoxon) in Appendix C. The results of the permutation test corroborate the significance of hypotheses, as with the exception of GloVe 50, only single hypotheses testing insignificant, regardless of modality.

[22] An intuitive animated explanation of a non-paired test can be found here: `https://www.jwilber.me/permutationtest/`. The difference for the paired test is that each data point in one sample is paired with another data point in the other

6.7 Informal correlation analyses

6.7.1 Between datasets and modalities

To get a better understanding how rankings differ between datasets, I informally assess Pearson's r for every possible pair (without claim to statistical significance), measuring linear correlation. Figure 11 shows a heatmap of correlation coefficients, with redder values indicating higher correlation. Correlation values between 0.4 and 0.6 are considered fair to strong, with higher values being moderate to very strong, depending on the scale (see [Akoglu, 2018]).

As expected from the box plots shown above, strong to very strong correlation is present among eye-tracking and EEG datasets, suggesting robustness on the modality level. Correlation between EEG and fMRI is non-existent, which is expected given the the issues described above. There is also no correlation between EEG and eye-tracking datasets, which is in line with the clear differences in ranking observed above and reinforces the intuition that the modalities measure different aspects. However I note that some correlation is observed between eye-tracking and fMRI as well asl EEG when BERT is excluded (not shown). Between EEG and fMRI, Pereira (small) also correlates noticeably with ZuCo datasets. I omit a discussion of fMRI correlation because of the issues decribed above.

6.7.2 Between regressions MSEs and extrinsic results

Following, I relate results aggregated on the modality level with Perone et al. [2018]'s findings regarding **downstream classification tasks**, **semantic relatedness and textual similarity tasks** and **linguistic probing tasks**. Image and caption retrieval tasks are excluded. Given the number of tasks, it is not possible to summarize them at this point and I refer to Appendix B for a brief overview for each tasks, adopted unchanged from the authors.

When treating USE variants separately, only seven embeddings match the author's evaluation. This is because the evaluation predates the publication of BERT and GloVe 50 cannot be fairly compared to the 300-dimensional GloVe embeddings tested by the authors. Due to this small overlap, an observed correlation on the task-level may be coincidental (see e.g Aggarwal and Ranganathan [2016]). Furthermore, InferSent and fastText cannot be exactly matched. For InferSent, the authors use the GloVe-based variant[23], whereas the fastText-based model is used here (see above).

[23]Correspondence with authors

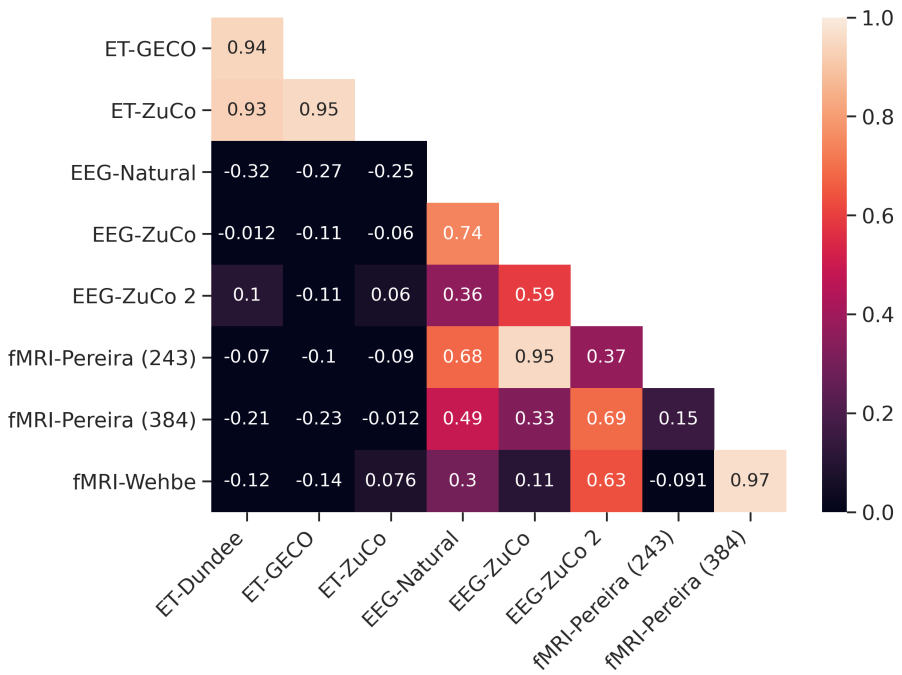


Figure 11: Heatmap of Pearsons's r between datasets across all modalities. Correlation values range from -1 to 1 (fully inverted to perfect correlation), but the heatmap color gradient is limited to 0 to 1, to highlight positive correlation

For fastText, the authors rely on the CC variant (subwords are implied), while the best-performing model out of five was chosen per dataset in the present evaluation, and marked differences were observed for EEG datasets. For ELMo, the 5.5B model was selected in almost all instances in the present evaluation, except for one eye-tracking source[24], which allows a nearly perfect match match. I emphasize that I make an exception for USE Transformer, which differs not only in training data or mode, but relies on a different architecture than USE DAN and as such, is a separate approach. Furthermore, it yields higher errors than USE DAN except for two fMRI datasets, yet performs better than USE DAN in many and much better in some downstream tasks across all three categories.

Due to these caveats, the relationship between the two sets of results cannot be formally established and tested for significance. However, under the (unproven) assumption that differences between approaches generally trump differences between variations of a particular approach, I argue that a rough estimate of rank correlation on the task level may at least serve as a starting point for subsequent evaluations, again without claim to statistical significance. As such, I calculate Spearman's ρ between the inverse of the MSE aggregated on the modality level and the corre-

[24] I note that the 5.5B model only improves one of four features, which however dominates the other features in scale and thus reduces the overall error.

sponding task measure.

Figure 12 shows per-modality heatmaps for each of the three task categories, with Pereira (small) shown separately, given the previously described issues with the other datasets. Eye-tracking results shows overall weak to moderate correlation with many downstream classification results and moderate to strong correlation with linguistic probing tasks. As both task categories are classification tasks, this suggests that some information leveraged from the embeddings is also used when learning to predict sequences of eye-tracking features. On the other hand, EEG correlates moderately to strongly with many semantic relatedness, which indicates that in predicting dimensionality-reduced concatenations of word-level EEG signals, a similar task is learnt. Given that the rankings between the modalities appear semi-reversed for both the classification and semantic relatedness task, this offers at least the prospect that eye-tracking and EEG datasets provide complementary information. If that is the case, the modalities may inform model selection and benefit cognitively informed NLP approaches when used in tandem. Finally, no (positive) correlation is observed for the fMRI modality, which is to be expected given that two of three datasets provide little to no discriminating information. However, even when only considering Pereira (small), for which the largest proportional differences between ranks were observed, no clear tendency arises. Hence, in the present evaluation setting, fMRI data is not predictive of extrinsic evaluation results.

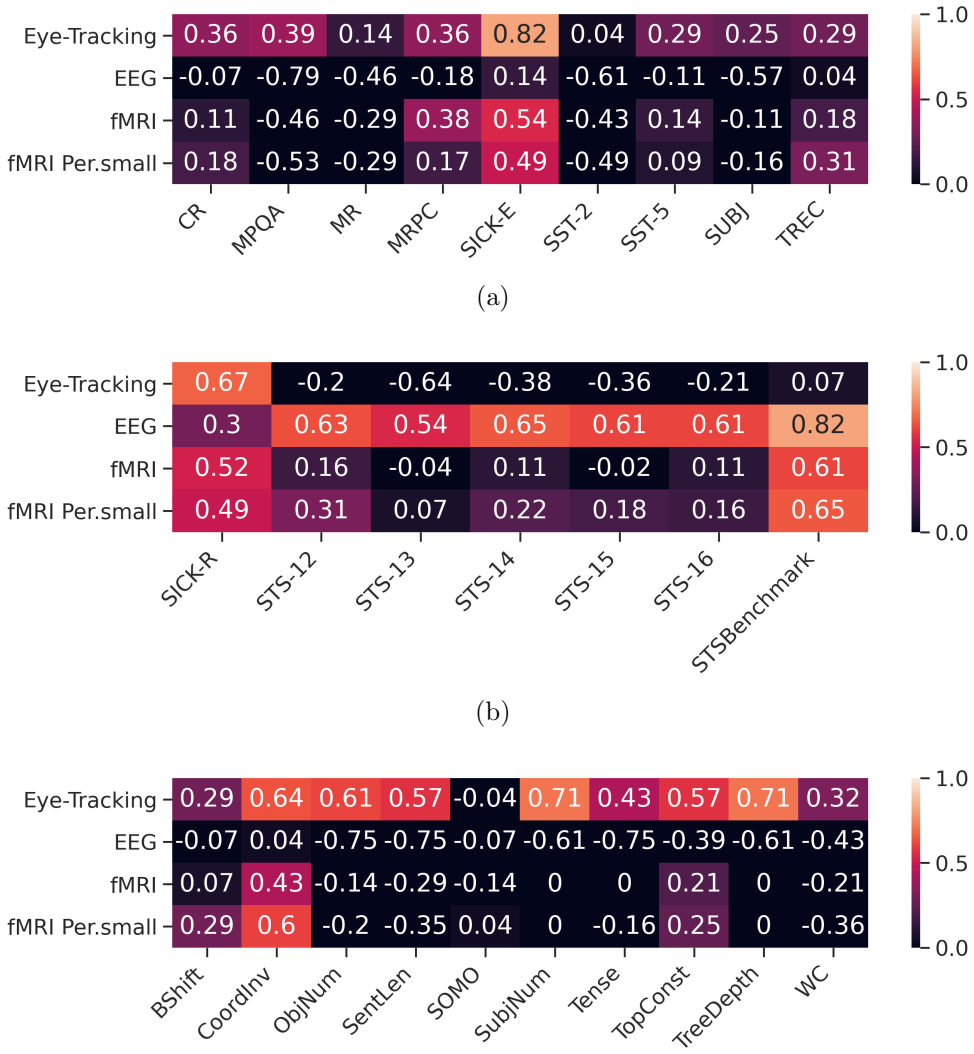


Figure 12: Heatmap of inverse MSEs of embeddings per modality, correlated with results of **downstream classification tasks** (a), **semantic relatedness and textual similarity tasks** (b), as well as **linguistic probing tasks** (c). Correlation values range from -1 to 1 (fully inverted to perfect correlation), but the heatmap color gradient is limited to 0 to 1, to highlight positive correlation.

7 Discussion

The observed performance of embeddings is necessarily influenced to some extent by differences in the size and content of corpora that are the basis of the pre-trained models, the pre-processing applied to those corpora, as well as tokenization during test time, with some approaches requiring external tokenization (Power-Mean, ELMo and averaging sentence baselines) while others use built-in tokenization methods of varying complexity. Furthermore, I agree with [Rücklé et al., 2018] that the difference in dimensionality is a confounding factor when evaluating more complex approaches against baselines. However, several observations relativize these considerations. In the case of fastText, I find that models trained on substantially more data or including subwords only show marked difference in EEG experiments. ELMo, a more sophisticated approach which [Perone et al., 2018] find to perform very competitively across many tasks, uses less training data for either pre-trained model than the smallest corpus of any pre-trained fastText model. More so, InferSent's overall peak performance (both presently and in extrinsic results) is contrasted by the comparably small sizes of NLI corpora, whilst building on fastText embeddings not leveraging subwords (here, in line with the author's recommendation) or GloVe embeddings (Perone et al.)

Dimensionality is of two-fold importance in the present evaluation. On one hand, vectors can encode more information with increasing dimensions. On the other, when dimensionality exceeds the number of data points reserved for training, it is likely not all salient information can be leveraged. Both effects can lead to predictions that do not adequately reflect the potential of the tested approach. However, in the present evaluation, the difference in dimensionality between the three averaging baselines GloVe 50, fastText and Power-Mean (50, 300 and 3600), are not reflected proportionally in the results. This is despite the fact that GloVe 50 is only used to establish a lower bound and does not adequately represent the approach. In addition, the 512-dimensional USE encoder matches or outperforms any of the three methods across all datasets. These observations relativize the importance of dimensionality in the present evaluation.

In contrast, a low ratio between data points and features is likely to significantly

contribute to the error boundary observed for the fMRI datasets (although the lower spread observed for the larger Pereira and Wehbe dataset also indicates noise issue). However, for the remaining datasets, there is no indication that approaches are gravely misrepresented due to data sparsity, given the low errors obtained by InferSent and Skip-Thought, which also generate the highest-dimensional vectors in the present evaluation (4096 and 4800 dimension, respectively).

Given these observations, a systematic bias confounding the present findings appears unlikely. However, at least in the case of Power-Mean, ELMo and BERT, some specific issues may be present (the first of which is not evaluated [Perone et al., 2018]). ELMo and BERT constitute anomalies, performing (very) competitively in extrinsic evaluations, yet obtaining only a fair ranking for eye-tracking and falling behind in various ways for EEG, as well as fMRI (ELMo only). This contrasts with findings by Hollenstein et al. [2019], where the representation yielded by the context-insensitive layer of ELMo obtains (relatively) low MSE values for several datasets across modalities. For BERT, it is possible that a fine-tuned variant would fare better. However, this would contradict the task-indepent approach proposed by [Søgaard, 2016]. On the other hand, the sentence representations I evaluated for these approaches all rely on some form of averaging, and ELMo and Power-Mean vectors also rely on concatenation, the latter inherently. This could indicate an issue with representations that are composites of separately encoded vectors. Furthermore, it is also possible that the presently used MLP regressor with two hidden layers has insufficient representational power to learn complex features of these approaches, such as the contextual features of ELMo. Given that Skip-Thought and InferSent also encode context, this would indicate that directly encoded, flat sentence representations are more easily learnt. I note that in the evaluation performed by [Perone et al., 2018], classification and linguistic probing tasks also relied on an MLP model with only a single hidden layer, a fixed dimensionality of 50 and no dropout, yet produced competitive results for ELMo. However, in the present evalation, the model predicts continuous and relatively high dimensional representations, compared with a finite number of discrete labels learnt for classification tasks. Lastly, it is also possible that data sparsity affects approaches to varying degrees and a change in ranking would be observed when training on larger datasets.

8 Conclusion

To conclude, I answer the research questions stated at the beginning of this thesis:

1. Sentence-level cognitive signals can be predicted by sentence embedding, obtaining low regression MSE and AED scores. For the tested eye-tracking and EEG datasets, AED scores of baselines do not differ in magnitude from embeddings, and for EEG and fMRI, proportional differences in MSE between embeddings tend to be small. However, in nearly all instances, results are significantly better than randomly generated baselines, as determined using a one-sided paired permutation test.

2. & 3. The relationship between the MSE and other evaluation results is not conclusively established. In an informal assessment of correlation, eye-tracking results show overall large correlation with linguistic probing tasks (another type of intrinsic evaluation), while correlation with downstream classification tasks are at best medium. The strongest correlations are observed between EEG and tasks covering semantic relatedness and textual similarity. For fMRI, no clear indication of correlation is present for any of the task categories, both on the aggregated level, as well as for the single dataset for which ranks are most dispersed.

4. [Perone et al., 2018] note that no single approach outperforms all others across all four evaluated task categories. Generally, InferSent, USE and ELMo show competitive overall performance. For InferSent, this is well reflected when predicting cognitive signals. Likewise, USE ranks highly for two of three modalities. However, it is represented by the DAN variant, which obtains similar errors to the transformer variant, while being substantially outperformed by the latter in many of the authors tasks across categories. This indicates that additional information encoded by the transformer variant is not leveraged in the present evaluation. Nonetheless, the performance of InferSent and USE substantiate the usefulness of comparably small but high-quality annotated datasets. Finally, ELMo and BERT present anomalies, performing overall unfavorably as well as inconsistently between datasets and modalities, despite

strong extrinsic performance observed particularly for BERT. This indicates limitations of the present evaluation, which could be adressed in various ways, such as different modes of pre-processing, alternative regression models or larger datasets.

The presented results substantiate previous findings that embeddings are predictive of cognitive signals and that differences between approaches meaningfully relate to the results obtained by other methods of assessing quality. As an initial cognitive assessment of sentence embeddings, I hope to provide a reference for subsequent evaluation of other recent and notable approaches which were presently omitted. As more cognitive datasets become available, future research may explore more fine-grained aspects such as the influence of text genre and the mode of consumption (reading or listening). Furthermore, evaluations of languages other than English are made possible by datasets such as the Russian Sentence Coprus [Laurinavichyute et al., 2019] and the Dutch Narrative Brain Dataset [Lopopolo et al., 2018]. This can also serve to evaluate multi-lingual approaches (e.g. Artetxe and Schwenk [2019]). In particular, combined datasets such as the ZuCo corpora allow to eliminate important confounding variables such as differences in text, subjects and experimental settings, further increasing the robustness of observed results. If larger datasets for EEG and fMRI are curated, it is likely that more information can be leveraged from high-dimensional embeddings, as this would alleviate the currently problematic ratio between the number of features and data points available for training. Finally, examining the influence of training data, model parameters, as well as the predictiveness of specific output layers (where applicable) may help to gain a more fine-grained understanding of individual approaches, benefitting research and practical applications alike.

Glossary

multi-layer perceptron Feedforward neural network with at least one densely connected hidden layer using a nonlinear activation function.

neural regression Non-linear regression relying on a neutral network to predict continuous scalars or vectors (as opposed to discrete values, as in classification).

References

S. Abnar, R. Ahmed, M. Mijnheer, and W. Zuidema. Experiential, distributional and dependency-based word embeddings have complementary roles in decoding brain activity. *arXiv preprint arXiv:1711.09285*, 2017.

R. Aggarwal and P. Ranganathan. Common pitfalls in statistical analysis: The use of correlation techniques. *Perspectives in clinical research*, 7(4):187, 2016.

E. Agirre, C. Banea, D. Cer, M. Diab, A. Gonzalez Agirre, R. Mihalcea, G. Rigau Claramunt, and J. Wiebe. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval-2016. 10th International Workshop on Semantic Evaluation*, pages 497–511, 2016.

H. Akoglu. User's guide to correlation coefficients. *Turkish journal of emergency medicine*, 18(3):91–93, 2018.

M. Artetxe and H. Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 2019.

D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

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

P. Baldi and P. J. Sadowski. Understanding dropout. In *Advances in neural information processing systems*, pages 2814–2822, 2013.

M. Barrett, Ž. Agić, and A. Søgaard. The dundee treebank. In *The 14th International Workshop on Treebanks and Linguistic Theories (TLT 14)*, 2015.

M. Barrett, J. Bingel, F. Keller, and A. Søgaard. Weakly supervised part-of-speech tagging using eye-tracking data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 579–584, 2016.

L. Beinborn, S. Abnar, and R. Choenni. Robust evaluation of language-brain encoding experiments. *International Journal of Computational Linguistics and Applications*, pages to–appear, 2019.

P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

S. Bowman, G. Angeli, C. Potts, and C. D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, 2015.

J. R. Brennan, E. P. Stabler, S. E. Van Wagenen, W.-M. Luh, and J. T. Hale. Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and language*, 157:81–94, 2016.

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

M. Catani, D. K. Jones, and D. H. Ffytche. 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, 2005.

D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*, 2017.

D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

B. Chiu, A. Korhonen, and S. Pyysalo. 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, 2016.

J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

M. Coltheart, B. Curtis, P. Atkins, and M. Haller. Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological review*, 100(4):589, 1993.

A. Conneau and D. Kiela. Senteval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, 2017.

A. Conneau, G. Kruszewski, G. Lample, L. Barrault, and M. Baroni. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, 2018.

U. Cop, N. Dirix, D. Drieghe, and W. Duyck. Presenting geco: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior research methods*, 49(2):602–615, 2017.

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

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

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

A. Dietrich and R. Kanso. A review of eeg, erp, and neuroimaging studies of creativity and insight. *Psychological bulletin*, 136(5):822, 2010.

B. Dolan, C. Quirk, and C. Brockett. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 350–356, 2004.

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

R. Dror, G. Baumer, S. Shlomov, and R. Reichart. 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, 2018.

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

S. L. Frank, L. J. Otten, G. Galli, and G. Vigliocco. Word surprisal predicts n400 amplitude during reading. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 878–883, 2013.

J. Gauthier and R. Levy. Linking artificial and human neural representations of language. *arXiv preprint arXiv:1910.01244*, 2019.

A. Gladkova and A. Drozd. 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, 2016.

Y. Goldberg. Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10(1):1–309, 2017.

J. Goodman. Classes for fast maximum entropy training. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings*, volume 1, pages 561–564. IEEE, 2001.

G. H. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities. By GH Hardy, JE Littlewood, G. Pólya..* University Press, 1952.

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

M. Henderson, R. Al-Rfou, B. Strope, Y.-H. Sung, L. Lukács, R. Guo, S. Kumar, B. Miklos, and R. Kurzweil. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*, 2017.

F. Hill, R. Reichart, and A. Korhonen. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695, 2015.

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

N. Hollenstein, J. Rotsztejn, M. Troendle, A. Pedroni, C. Zhang, and N. Langer. Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. *Scientific data*, 5(1):1–13, 2018.

N. Hollenstein, A. de la Torre, N. Langer, and C. Zhang. Cognival: A framework for cognitive word embedding evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 538–549, 2019.

N. Hollenstein, M. Troendle, C. Zhang, and N. Langer. 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, 2020.

M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, 2004.

E. B. Huey. *The psychology and pedagogy of reading.* The Macmillan Company, 1908.

A. G. Huth, W. A. De Heer, T. L. Griffiths, F. E. Theunissen, and J. L. Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, 2016.

M. Iyyer, V. Manjunatha, J. Boyd-Graber, and H. Daumé III. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1681–1691, 2015.

A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, 2017.

A. Kennedy, R. Hill, and J. Pynte. The dundee corpus. In *Proceedings of the 12th European conference on eye movement*, 2003.

T. Kenter, A. Borisov, and M. De Rijke. Siamese cbow: Optimizing word embeddings for sentence representations. *arXiv preprint arXiv:1606.04640*, 2016.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302, 2015.

S. Klerke, Y. Goldberg, and A. Søgaard. Improving sentence compression by learning to predict gaze. *arXiv preprint arXiv:1604.03357*, 2016.

M. Kutas and K. D. Federmeier. Thirty years and counting: finding meaning in the n400 component of the event-related brain potential (erp). *Annual review of psychology*, 62:621–647, 2011.

A. K. Laurinavichyute, I. A. Sekerina, S. Alexeeva, K. Bagdasaryan, and R. Kliegl. Russian sentence corpus: Benchmark measures of eye movements in reading in russian. *Behavior research methods*, 51(3):1161–1178, 2019.

Y. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller. Efficient backprop. In *Neural Networks: Tricks of the Trade*, pages 9–50. Springer, 1998.

X. Li and D. Roth. Learning question classifiers. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7, 2002.

D. J. Liebling and S. Preibusch. Privacy considerations for a pervasive eye tracking world. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, pages 1169–1177, 2014.

T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755, 2014.

A. Lopopolo, S. Frank, A. van den Bosch, A. Nijhof, and R. Willems. The narrative brain dataset (nbd), an fmri dataset for the study of natural language processing in the brain. In *Devereux, B.; Shutova, E.; Huang, C.-R.(ed.), Proceedings of LREC 2018 Workshop" Linguistic and Neuro-Cognitive Resources (LiNCR)"*, pages 8–11, 2018.

C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014.

M. Marelli, S. Menini, M. Baroni, L. Bentivogli, R. Bernardi, R. Zamparelli, et al. A sick cure for the evaluation of compositional distributional semantic models. In *LREC*, pages 216–223, 2014.

F. M. Miezin, L. Maccotta, J. Ollinger, S. Petersen, and R. Buckner. 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, 2000.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems 26*, pages 3111–3119, 2013b.

A. Mishra and P. Bhattacharyya. *Cognitively Inspired Natural Language Processing: An Investigation Based on Eye-tracking.* Springer, 2018.

T. M. Mitchell, S. V. Shinkareva, A. Carlson, K.-M. Chang, V. L. Malave, R. A. Mason, and M. A. Just. Predicting human brain activity associated with the meanings of nouns. *science*, 320(5880):1191–1195, 2008.

N. Mrkšić, I. Vulić, D. Ó Séaghdha, I. Leviant, R. Reichart, M. Gašić, A. Korhonen, and S. Young. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the Association for Computational Linguistics*, 5:309–324, 2017.

B. Murphy, L. Wehbe, and A. Fyshe. Decoding language from the brain. *Language, cognition, and computational models*, pages 53–80, 2018.

E. Niedermeyer and F. L. da Silva. *Electroencephalography: basic principles, clinical applications, and related fields.* Lippincott Williams & Wilkins, 2004.

B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *arXiv preprint cs/0409058*, 2004.

B. Pang and L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *arXiv preprint cs/0506075*, 2005.

R. Parker, D. Graff, J. Kong, K. Chen, and K. Maeda. English gigaword fifth edition ldc2011t07 (tech. rep.). Technical report, Technical Report. Linguistic Data Consortium, Philadelphia, 2011.

J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

F. Pereira, B. Lou, B. Pritchett, S. Ritter, S. J. Gershman, N. Kanwisher, M. Botvinick, and E. Fedorenko. Toward a universal decoder of linguistic meaning from brain activation. *Nature communications*, 9(1):1–13, 2018.

C. S. Perone, R. Silveira, and T. S. Paula. Evaluation of sentence embeddings in downstream and linguistic probing tasks. *arXiv preprint arXiv:1806.06259*, 2018.

M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In *Proc. of NAACL*, 2018.

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

K. Rayner and S. A. Duffy. Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & cognition*, 14(3):191–201, 1986.

F. Rosenblatt. Principles of neurodynamics: perceptrons and the theory of brain mechanisms' to van der malsburg c.(1986) frank rosenblatt: Principles of neurodynamics: Perceptrons and the theory of brain mechanisms. *Brain Theory. Berlin, Heidelberg: Springer*, 1962.

J. Rowling. *Harry Potter and the Sorcerer's Stone*. Scholastic Press, 1999.

A. Rücklé, S. Eger, M. Peyrard, and I. Gurevych. Concatenated power mean word embeddings as universal cross-lingual sentence representations. *arXiv*, 2018.

R. Salmelin. Clinical neurophysiology of language: the meg approach. *Clinical Neurophysiology*, 118(2):237–254, 2007.

B. Schölkopf, A. Smola, and K.-R. Müller. Kernel principal component analysis. In *International conference on artificial neural networks*, pages 583–588, 1997.

D. Schwartz and T. Mitchell. Understanding language-elicited eeg data by predicting it from a fine-tuned language model. *arXiv preprint arXiv:1904.01548*, 2019.

R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. 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, 2013.

A. Søgaard. Evaluating word embeddings with fmri and eye-tracking. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 116–121, 2016.

J. Steil, I. Hagestedt, M. X. Huang, and A. Bulling. Privacy-aware eye tracking using differential privacy. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, pages 1–9, 2019.

Y. Sun, S. Wang, Y.-K. Li, S. Feng, H. Tian, H. Wu, and H. Wang. Ernie 2.0: A continual pre-training framework for language understanding. In *AAAI*, pages 8968–8975, 2020.

I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems NIPS*, pages 3104–3112, 2014.

W. L. Taylor. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433, 1953.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *31st Conference on Neural Information Processing Systems (NIPS 2017*, pages 5998–6008, 2017.

T. Von der Malsburg and S. Vasishth. What is the scanpath signature of syntactic reanalysis? *Journal of Memory and Language*, 65(2):109–127, 2011.

E. M. Voorhees and D. M. Tice. Building a question answering test collection. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 200–207, 2000.

I. Vulić, N. Mrkšić, R. Reichart, D. Ó Séaghdha, S. Young, and A. Korhonen. Morph-fitting: Fine-tuning word vector spaces with simple language-specific rules. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 56–68, Vancouver, Canada, July 2017.

S. I. Wang and C. D. Manning. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 90–94, 2012.

L. Wehbe, B. Murphy, P. Talukdar, A. Fyshe, A. Ramdas, and T. Mitchell. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one*, 9(11), 2014a.

L. Wehbe, A. Vaswani, K. Knight, and T. Mitchell. 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, 2014b.

K. Weinberger, A. Dasgupta, J. Langford, A. Smola, and J. Attenberg. Feature hashing for large scale multitask learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 1113–1120, 2009.

J. Wiebe, T. Wilson, and C. Cardie. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210, 2005.

Wikipedia. Electroencephalography — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=Electroencephalography&oldid=959739114`, 2020. [Online; accessed 01-June-2020].

A. Williams, N. Nangia, and S. R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1112–1122, 2018.

S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the 2015 IEEE international conference on computer vision*, pages 19–27, 2015.

A A command-line tool for large-scale neural evaluation of word and sentence embeddings and cognitive sources

The command-line tool[1] developed in the context of this thesis offers functions to download a variety of word and sentence embeddings, as well as generate baseline sentence embeddings from word embeddings. For complex embeddings, word and sentence representations are dynamically generated from the vocabulary and sentence listings obtained from cognitive sources (datasets), and can be updated to reflect newly added sources. Users may also import externally generated sentence and word-level vectors. For every embedding, a set of random baselines is generated with fixed non-linearly increasing seeds, for the purpose of significance testing. By default, ten such baselines are generated. All word- and sentence-level cognitive sources evaluated in this work and Hollenstein et al. [2019] can be downloaded and imported with a single command. Word-level sources are available with fixed scaling and pre-sampled (fMRI only), whereas sentence-level sources are provided unscaled and unsampled and are transformed on-the-fly within the cross-validation loops.

Experiments are defined as pairs of cognitive sources and embeddings (as well as random baselines, if associated), where the cognitive source constitutes the parent entity, with embeddings assigned to it. The tool offers fine-grained control over modifying grid search parameters across multiple embeddings, as well as cognitive data sources or entire modalities. Each modification triggers the creation of a configuration backup, allowing for convenient tracing of changes and restoring previous parametrizations. A variety of convenience commands exist for managing and viewing properties of configurations, embeddings and cognitive source.

Experimental runs are carried out in parallel, with results being continuously read

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

and post-processed for subsequent significance testing, aggregation and reporting. For embeddings with identical parametrization (dimensionality and hyper-parameter grid), random embedding results can be reused to reduce overall runtime.

Significance testing relies on the Wilcoxon signed-rank test implementation provided by the `scipy`[2] library and (one-sided) paired permutation test implementation provided by the `permute`[3] package. Significance test results, as well as aggregated MSE and significance ratios can be directly viewed in the shell. Alternatively, a comprehensive, interactive HTML report can be generated[4]

The HTML report contains sortable tables of aggregated results of embeddings per modality (MSEs and significance ratios), as well as detailed results for individual cognitive sources with respect to embeddings and random baselines, with the option to aggregate subjects or features if present. Optionally, searchable, dynamic tables may be added, showing the unit-level error per word for each sentence or word evaluated for a cognitive source (and depending on the modality, feature or subject). When including random baselines and performing multiple runs, history plots give a convenient overview of the change of aggregated statistics such as average MSE and significance ratio. Finally, training history plots may be added for a quick and immediate overview over convergence times and potential overfitting issues.

[2] `https://docs.scipy.org/doc/scipy/reference/`

[3] `https://github.com/statlab/permute`

[4] While primarily intended for viewing in a GUI browser, the representation is compatible with text-mode browser such as `links` and `lynx`. When a command-line browser is installed and registered as the default browser, specifying to open the HTML will directly view the report in the terminal after generating. This allows for immediate viewing of results on a remote compute server accessed via SSH.

B Overview of evaluation tasks by Perone et al. [2018]

For convenience, the following tables briefly summarize the tasks of three of four categories evaluated by Perone et al. [2018], as considered for correlation analysis in Section 6.7.2: Downstream classification, semantic relatedness and linguistic probing. Tables are taken directly from [Perone et al., 2018], with some formatting adjustments. Please refer to the original article for citations to each dataset.

	Dataset	**Task**	**Example**	**Output**
CR	Customer Reviews [Hu and Liu, 2004]	Sentiment analysis of customer	We tried it out Christmas night and it worked great .	Positive
MPQA	Multi-Perspective Question and Answering [Wiebe et al., 2005]	Evaluation of opinion polarity	Don't want	Negative
MR	Movie Reviews [Pang and Lee, 2005]	Sentiment analysis of movie reviews	Too slow for a younger crowd, too shallow for an older one .	Negative
SST-2	Stanford Sentiment Analysis 2 [Socher et al., 2013]	Sentiment analysis with two classes: Negative and Positive	Audrey Tautou has a knack for picking roles that magnify her [..]	Positive
SST-5	Stanford Sentiment Analysis 5 [Socher et al., 2013]	Sentiment analysis with 5 classes that range from 0 (most negative) to 5 (most positive)	Nothing about this movie works	0
SUBJ	Subjectivity / Objectivity [Pang and Lee, 2004]	Classify the sentence as Subjective or Objective	A movie that doesn't aim too high , but doesn't need to .	Subjective
TREC	Text REtrieval Conference [Voorhees and Tice, 2000]	Question and answering	What are the twin cities ?	LOC:city

Table 7: Downstream classification tasks description and samples.

	Dataset	Task	Example	Output
COCO	Common Objects in Context [Lin et al., 2014]	Image-caption retrieval (ICR)	-	A group of people on some horses riding through the beach
MRPC	Microsoft Research Paraphrase Corpus [Dolan et al., 2004]	Classify whether a pair of sentences capture a paraphrase relationship	The procedure is generally performed in the scond or third trimester	The technique is used during the scond and, occasionally, third trimester of pregnancy
STS	Semantic Text Similarity [Cer et al., 2017]	To measure the semantic similarity between two sentences from 0 (not similar) to 5 (very similar)	Liquid ammonia leak kills 15 in Shanghai	Liquid ammonia leak kills at least 15 in Shanghai
SICK-E	Sentences Involving Compositional Knowledge - Entailment [Marelli et al., 2014]	To measure semantics in terms of Entailment, Contradiction, or Neutral	A man is sitting on a chair and rubbing his heyes	There is no man sitting on a chair and rubbing his eyes
SICK-R	Sentences Involving Compositional Knowledge - Relatedness [Marelli et al., 2014]	To measure the degree of semantic relatedness between sentences from 0 (not related) to 5 (related)	A man is singing a song and playing the guitar	A man is opening a package that contains headphones
SNLI	Stanford Natural Language Inference [Bowman et al., 2015]	To measure semantics in terms of Entailment, Contradiction, or Neutral	A small girl wearing a pink jacked is riding on a carousel	The carousel is moving

Table 8: Downstream semantic relatedness and textual similarity tasks descriptions and samples.

	Task	Task	Example	Output
BShift	Bigram Shift	Whether two words (tokens) in a sentence have been inverted	This is my Eve Christmas .	Inverted
CoordInv	Coordination Inversion	Sentences comprised of two coordinate clauses. Detect whether clauses are inverted	I returned to my work, and Lisa headed for her office .	Inverted
ObjNum	Object Number	Number of the direct object in the main clause (singular and plural)	He received the 200 points .	NNS (Plural)
SentLen	Sentence Length	Predict the sentence length among 6 classes, which are length intervals	I can't wait to show you and Mr. Taylor .	9 - 12 words
SOMO	Semantic Odd Man Out	Random noun or verb replaced in the sentence by another noun or verb. Detect whether the sentence has been modified	Tomas surmised as well .	Changed
SubjNum	Subject Number	Number of the subject in the main clause (singular and plural)	If there was ever a time to let loose , this vacation would have to be it .	Singular
Tense	Past Present	Whether the main verb in the sentence is in the past or present tense	She smiled at him , her eye alight with love .	Present
TopConst	Top-Constituent	Classification task, where the classes are given by the 19 most common top-constituent sequences in the corpus	Did he buy anything from Troy ?	VBD_NP_VP_
TreeDepth	Depth of Syntactic Tree	Predict the maximum depth of the syntactic tree of the sentence	The leaves were in various of stages of life .	10
WC	Word Content	Predict which of the target words (among 1000) appear in the sentence	She eyed him skeptically .	eyed

Table 9: Linguistic probing tasks description and samples [Conneau et al., 2018]

C Tables

For all tables showing error values, lowest (best) values are formatted bold. For tables 10 to 12, this corresponds to the column minimum and to the row minimum for all other tables.

Tables 10 to 12 show the average MSE values per dataset and approach/corresponding baseline. For fMRI, MSE values are averaged among subjects and for eye-tracking, averaging was performed across features. The tables complement the box plots in chapter 6, which visualize median, spread and extrema of error values. Note that the usefulness of the averaged MSE value is limited for embeddings and datasets with large inter-quantile ranges (IQRs).

Note: Tables 29 onward are sorted alphabetically along both axis.

	Dundee		**GECO**		**ZuCo**	
	Baseline	**Embed.**	**Baseline**	**Embed.**	**Baseline**	**Embed.**
GloVe (50)	**0.023828**	0.022485	**0.009411**	0.007934	0.019646	0.019532
fastText	0.024958	0.021470	0.009602	0.008360	0.020103	0.018732
Power-Mean	0.026367	0.019169	0.010528	0.006399	0.021397	0.015941
ELMo	0.026335	0.016187	0.011704	0.005789	0.021421	0.016088
BERT	0.026819	**0.015772**	0.010747	0.005652	0.021850	0.014821
Skip-Thought	0.026618	0.015820	0.010694	**0.005418**	0.021515	0.014920
InferSent	0.026580	0.015807	0.011042	0.005706	0.021306	**0.014484**
USE	0.026081	0.018470	0.009938	0.009938	0.020679	0.017636

Table 10: Average regression MSE for eye-tracking experiments, averaged across five folds and five (Dundee) and four (GECO, ZuCo) eye-tracking features respectively.

Tables 14 to 22 show the MAE for the (non-exclusive) proportions of sentences containing at least one occurrence of a specific linguistic feature. All features have been obtained through the largest currently available `spaCy` model for English[1]. Features and corresponding labels are: 1) *grammatical dependencies* adjectival clause (acl),

[1] `https://spacy.io/models/en#en_core_web_lg`

	Natural Speech		ZuCo		ZuCo 2	
	Baseline	**Embed.**	**Baseline**	**Embed.**	**Baseline**	**Embed.**
GloVe (50)	0.000458	0.000163	0.000300	0.000036	0.000458	0.000110
fastText	**0.000183**	0.000169	0.000100	0.000028	0.002153	**0.000094**
Power-Mean	0.000258	0.000164	0.000136	0.000031	0.000383	0.000112
ELMo	0.000198	0.000173	0.000071	0.000031	0.000226	0.000114
BERT	0.000193	0.000181	0.000055	0.000044	0.000151	0.000106
Skip-Thought	0.000305	0.000166	0.000167	0.000027	0.000743	0.000095
InferSent	0.000492	**0.000160**	0.000064	0.000026	0.001105	**0.000094**
USE	0.000186	0.000161	**0.000047**	**0.000024**	0.000465	**0.000094**

Table 11: Average regression MSE for EEG experiments, averaged across five folds

	Pereira (243)		Pereira (384)		Wehbe	
	Baseline	**Embed.**	**Baseline**	**Embed.**	**Baseline**	**Embed.**
GloVe (50)	0.000579	0.001018	**0.000830**	0.000769	0.010000	0.009877
fastText	**0.000415**	0.000384	0.001010	**0.000747**	0.010263	0.009838
Power-Mean	0.000942	0.000513	0.000798	0.000750	**0.009867**	0.009841
BERT	0.000571	0.001825	0.001026	0.000769	0.010079	0.009839
ELMo	0.000597	0.000438	0.000880	0.000825	0.010255	0.010002
Skip-Thought	0.000504	**0.000371**	0.000925	**0.000747**	0.010891	**0.009833**
InferSent	0.000517	**0.000371**	0.000905	**0.000747**	0.010965	0.009834
USE	0.001117	0.000375	0.001018	0.000749	0.010957	0.009837

Table 12: Average regression MSE for fMRI experiments, averaged across five folds and five (Pereira small) and eight (Pereira large, Wehbe) subjects respectively.

compound, clausal subject (csubj), negation (neg), preprositional modifier (prepr) and relative clause (relcl); 2) *named entity types* geopolitical entity (GPE), location (LOC), nationalities and religious/political group (NORP), organisation (ORG) and person (PER), occurrence of three or more (proper) nouns (POS 3+ NN), combined occurrence of a (proper) noun, adjective and main verb (POS NN & J & V), as well as three-way sentiment. Features occurring in less than ten sentences are omitted. Power-Mean and Skip-Thought are abbreviated as PM and ST, respectively.

Tables 29 to 33 show the final parameter grid relating to batch size (B) and layer size (L) (other parameters remain fixed, with the exception of number of epochs, which was reduced from 100 to 50, with the latter being sufficient to obtain convergence. See section TODO). Layer sizes correspond to both hidden layers used in the model specification. Eye-tracking features are abbreviated as follows.

	Eye-Tracking	EEG	fMRI
GloVe	10/13	**3/3**	**21/21**
fastText	12/13	**3/3**	**21/21**
ELMo	**13/13**	**3/3**	**21/21**
BERT	**13/13**	**3/3**	**21/21**
Skip-Thought	**13/13**	**3/3**	**21/21**
InferSent	**13/13**	**3/3**	**21/21**
Power-Mean	**13/13**	**3/3**	**21/21**
USE	**13/13**	**3/3**	**21/21**

Table 13: Significance ratios of hypotheses per modality, as obtained through the two-tailed Wilcoxon signed-rank test. $\alpha = 0.01$ with Bonferroni correction per modality.

FFD: first fixation duration, **GD**: gaze duration, **FP**: fixation probability, **MFD**: mean fixation duration, **TFD**: total fixation duration, **TRT**: total reading time and **nFix**: number of fixations.

	count	BERT	ELMo	fastText	GloVe	Infersent	PM	ST	USE
Dep. acl	372	0.081	0.083	0.092	0.093	0.08	0.088	**0.078**	0.084
Dep. compound	1286	0.074	0.077	0.087	0.088	0.073	0.082	**0.071**	0.078
Dep. csubj	43	0.073	0.075	0.082	0.083	**0.073**	0.08	**0.073**	0.076
Dep. neg	388	0.069	0.071	0.081	0.083	0.067	0.078	**0.066**	0.073
Dep. prep	2017	0.069	0.072	0.081	0.083	0.067	0.077	**0.066**	0.073
Dep. relcl	610	0.082	0.084	0.093	0.093	0.081	0.089	**0.079**	0.085
NE GPE	295	0.079	0.081	0.091	0.091	0.077	0.086	**0.076**	0.082
NE LOC	65	0.087	0.088	0.099	0.1	0.086	0.094	**0.084**	0.09
NE NORP	247	0.083	0.085	0.095	0.095	0.082	0.09	**0.08**	0.087
NE ORG	419	0.079	0.082	0.092	0.092	0.077	0.086	**0.075**	0.082
NE PERSON	555	0.076	0.079	0.088	0.089	0.075	0.084	**0.074**	0.081
POS 3+ NN	811	0.082	0.084	0.093	0.093	0.081	0.088	**0.079**	0.085
POS NN & J & V	1864	0.07	0.073	0.083	0.084	0.068	0.078	**0.067**	0.074
Sentiment (neg.)	124	0.061	0.064	0.075	0.077	**0.058**	0.071	**0.058**	0.066
Sentiment (neut.)	1838	0.066	0.069	0.079	0.082	0.064	0.076	**0.063**	0.07
Sentiment (pos.)	404	0.066	0.068	0.077	0.079	0.063	0.074	**0.062**	0.069

Table 14: Mean of the absolute error averaged across dimensions (MAED), per ling. feature, for the eye-tracking dataset **Dundee**

	count	BERT	ELMo	fastText	GloVe	Infersent	PM	ST	USE
Dep. acl	198	0.058	0.058	0.066	0.065	0.061	0.061	**0.054**	0.061
Dep. compound	1117	0.045	0.046	0.054	0.052	0.046	0.048	**0.042**	0.049
Dep. csubj	22	0.064	0.062	0.072	0.068	0.065	0.064	**0.058**	0.066
Dep. neg	782	0.037	0.038	0.045	0.043	0.037	0.038	**0.034**	0.041
Dep. prep	2700	0.041	0.042	0.048	0.047	0.042	0.043	**0.038**	0.044
Dep. relcl	453	0.056	0.057	0.065	0.063	0.058	0.059	**0.052**	0.06
NE GPE	44	0.049	0.05	0.058	0.056	0.05	0.051	**0.046**	0.052
NE LOC	10	0.044	0.046	0.054	0.05	0.049	0.046	**0.043**	0.048
NE NORP	28	0.048	0.048	0.056	0.055	0.05	0.048	**0.044**	0.051
NE ORG	158	0.045	0.045	0.053	0.053	0.046	0.047	**0.042**	0.049
NE PERSON	1527	0.04	0.041	0.048	0.046	0.04	0.041	**0.037**	0.044
POS 3+ NN	321	0.061	0.062	0.071	0.069	0.064	0.066	**0.058**	0.066
POS NN & J & V	1891	0.044	0.045	0.052	0.051	0.045	0.046	**0.041**	0.047
Sentiment (neg.)	351	0.032	0.034	0.04	0.039	0.032	0.034	**0.03**	0.036
Sentiment (neut.)	4048	0.032	0.033	0.04	0.038	0.031	0.032	**0.029**	0.036
Sentiment (pos.)	674	0.035	0.037	0.043	0.041	0.035	0.036	**0.032**	0.039

Table 15: Mean of the absolute error averaged across dimensions (MAED), per ling. feature, for the eye-tracking dataset **GECO**

	count	BERT	ELMo	fastText	GloVe	Infersent	PM	ST	USE
Dep. acl	93	0.083	0.085	0.092	0.094	**0.08**	0.092	**0.08**	0.087
Dep. compound	448	0.072	0.076	0.081	0.083	**0.068**	0.078	0.07	0.077
Dep. neg	58	0.065	0.067	0.074	0.076	**0.062**	0.07	0.063	0.071
Dep. prep	599	0.068	0.071	0.077	0.079	**0.064**	0.074	0.066	0.073
Dep. relcl	160	0.081	0.084	0.089	0.091	**0.079**	0.088	0.08	0.086
NE GPE	153	0.075	0.081	0.085	0.087	**0.071**	0.082	0.074	0.081
NE LOC	12	0.097	0.104	0.114	0.115	**0.097**	0.107	0.099	0.111
NE NORP	69	0.077	0.08	0.088	0.09	**0.073**	0.085	0.077	0.085
NE ORG	182	0.076	0.081	0.085	0.088	**0.072**	0.081	0.074	0.082
NE PERSON	314	0.074	0.079	0.084	0.085	**0.071**	0.081	0.073	0.08
POS 3+ NN	227	0.081	0.084	0.089	0.09	**0.077**	0.087	0.079	0.085
POS NN & J & V	485	0.069	0.071	0.077	0.079	**0.065**	0.075	0.067	0.074
Sentiment (neg.)	44	0.056	0.056	0.065	0.07	**0.052**	0.06	0.053	0.061
Sentiment (neut.)	522	0.067	0.07	0.077	0.079	**0.063**	0.073	0.065	0.072
Sentiment (pos.)	134	0.065	0.067	0.074	0.078	**0.06**	0.072	0.062	0.07

Table 16: Mean of the absolute error averaged across dimensions (MAED), per ling. feature, for the eye-tracking dataset **ZuCo**

	count	BERT	ELMo	fastText	GloVe	Infersent	PM	ST	USE
Dep. acl	25	0.0059	0.0055	0.0043	0.0043	**0.0041**	0.0044	0.0046	**0.0041**
Dep. compound	113	0.0073	0.0069	0.006	0.0058	**0.0056**	0.0059	0.0061	0.0057
Dep. neg	84	0.0057	0.0052	0.0044	0.0045	**0.0041**	0.0044	0.0046	**0.0041**
Dep. prep	391	0.0062	0.0057	0.0048	0.0047	**0.0045**	0.0048	0.005	**0.0045**
Dep. relcl	118	0.0073	0.0069	0.0059	0.0059	**0.0057**	0.006	0.0062	**0.0057**
NE GPE	12	0.0046	0.0042	0.0032	0.0031	**0.0029**	0.0032	0.0035	0.003
NE LOC	10	0.0054	0.0052	0.0041	0.0042	**0.0039**	0.0042	0.0043	**0.0039**
NE PERSON	11	0.0048	0.0043	0.0034	0.0034	**0.0031**	0.0035	0.0037	0.0043
POS 3+ NN	21	0.0056	0.0053	0.0042	0.0041	**0.004**	0.0043	0.0045	**0.004**
POS NN & J & V	368	0.0061	0.0056	0.0047	0.0046	**0.0044**	0.0047	0.0049	**0.0044**
Sentiment (neg.)	27	0.0053	0.0048	0.0038	0.0039	**0.0036**	0.0039	0.0041	**0.0036**
Sentiment (neut.)	551	0.0056	0.005	0.0042	0.0041	**0.0038**	0.0042	0.0044	**0.0038**
Sentiment (pos.)	117	0.0052	0.0046	0.0039	0.0037	**0.0035**	0.0038	0.004	0.0036

Table 17: Mean of the absolute error averaged across dimensions (MAED), per ling. feature, for the EEG dataset **Natural Speech**

	count	BERT	ELMo	fastText	GloVe	Infersent	PM	ST	USE
Dep. acl	84	0.005	0.004	0.0034	0.0043	0.0033	0.0039	0.0035	**0.0032**
Dep. compound	448	0.0047	0.0035	0.0031	0.0036	0.0029	0.0035	0.003	**0.0028**
Dep. neg	58	0.005	0.004	0.0033	0.004	**0.0032**	0.0039	0.0033	**0.0032**
Dep. prep	594	0.0046	0.0035	0.003	0.0036	0.0028	0.0035	0.0029	**0.0027**
Dep. relcl	159	0.0049	0.0038	0.0033	0.004	**0.0031**	0.0037	0.0033	**0.0031**
NE GPE	153	0.0047	0.0034	0.003	0.0036	0.0029	0.0035	0.003	**0.0028**
NE LOC	13	0.0049	0.0037	0.0033	0.0037	0.0032	0.0039	0.0032	**0.0031**
NE NORP	69	0.0048	0.0035	0.0031	0.0037	0.0029	0.0035	0.0031	**0.0028**
NE ORG	176	0.0048	0.0036	0.0031	0.0038	0.0029	0.0036	0.003	**0.0028**
NE PERSON	314	0.0047	0.0035	0.0032	0.0037	0.0029	0.0036	0.003	**0.0028**
POS 3+ NN	226	0.0048	0.0037	0.0031	0.0038	0.003	0.0036	0.0031	**0.0029**
POS NN & J & V	484	0.0046	0.0035	0.0029	0.0036	0.0028	0.0034	0.0029	**0.0027**
Sentiment (neg.)	44	0.0044	0.0034	0.0027	0.0033	0.0026	0.0033	0.0027	**0.0025**
Sentiment (neut.)	522	0.0046	0.0034	0.003	0.0035	0.0028	0.0035	0.0029	**0.0027**
Sentiment (pos.)	134	0.0046	0.0035	0.0029	0.0036	0.0028	0.0035	0.0028	**0.0027**

Table 18: Mean of the absolute error averaged across dimensions (MAED), per ling. feature, for the EEG dataset **ZuCo**

label	count	BERT	ELMo	fastText	GloVe	Infersent	PM	ST	USE
Dep. acl	34	0.0067	0.0071	**0.006**	0.0066	**0.006**	0.0071	0.0061	**0.006**
Dep. compound	263	0.0067	0.007	**0.0059**	0.0067	**0.0059**	0.0071	0.006	**0.0059**
Dep. neg	17	0.0068	0.0073	**0.0061**	0.0071	**0.0061**	0.0075	0.0062	**0.0061**
Dep. prep	307	0.0067	0.007	**0.0059**	0.0067	**0.0059**	0.0071	0.006	**0.0059**
Dep. relcl	56	0.0069	0.0072	**0.0061**	0.0065	**0.0061**	0.0071	0.0062	**0.0061**
NE GPE	94	0.0066	0.0069	**0.0058**	0.0068	**0.0058**	0.007	0.0059	**0.0058**
NE LOC	11	0.0069	0.0072	**0.0061**	0.0067	0.0062	0.007	0.0063	0.0062
NE NORP	52	0.0069	0.0072	**0.0061**	0.007	0.0062	0.0074	0.0062	0.0062
NE ORG	103	0.0067	0.0071	**0.0059**	0.0067	**0.0059**	0.0071	0.006	**0.0059**
NE PERSON	228	0.0068	0.0071	**0.006**	0.0068	**0.006**	0.0071	0.0061	**0.006**
POS 3+ NN	133	0.0068	0.0071	**0.006**	0.0067	**0.006**	0.0071	0.0061	**0.006**
POS NN & J & V	222	0.0068	0.0071	**0.006**	0.0068	**0.006**	0.0072	0.0061	**0.006**
Sentiment (neut.)	301	0.0067	0.007	**0.0059**	0.0067	**0.0059**	0.007	0.006	**0.0059**
Sentiment (pos.)	36	0.0064	0.0069	**0.0058**	0.0066	**0.0058**	0.0071	0.0059	**0.0058**

Table 19: Mean of the absolute error averaged across dimensions (MAED), per ling. feature, for the EEG dataset **ZuCo 2**

	count	BERT	ELMo	fastText	GloVe	InferSent	PM	ST	USE
Dep. acl	12	**0.0092**	0.012	0.0099	0.0102	**0.0092**	0.0142	**0.0092**	**0.0092**
Dep. compound	92	0.0098	0.0123	0.0101	0.0112	**0.0094**	0.0144	**0.0094**	0.0095
Dep. neg	17	**0.0081**	0.0111	0.0088	0.0087	**0.0081**	0.0132	**0.0081**	**0.0081**
Dep. prep	192	0.0111	0.0124	0.0102	0.0116	**0.0095**	0.0145	**0.0095**	0.0096
Dep. relcl	27	0.0107	0.0124	0.0101	0.01	**0.0094**	0.0145	**0.0094**	**0.0094**
POS NN & J & V	180	0.0106	0.0122	0.01	0.0108	**0.0093**	0.0143	**0.0093**	0.0094
Sentiment (neg.)	16	**0.0094**	0.0122	0.0101	0.0131	**0.0094**	0.0144	**0.0094**	0.0095
Sentiment (neut.)	184	0.0099	0.0123	0.0101	0.0113	**0.0094**	0.0144	**0.0094**	0.0095
Sentiment (pos.)	43	0.0138	0.0116	0.0092	0.0095	**0.0086**	0.0137	**0.0086**	0.0088

Table 20: Mean of the absolute error averaged across dimensions (MAED), per ling. feature, for the fMRI dataset **Pereira (small)**

	count	BERT	ELMo	fastText	GloVe	InferSent	PM	ST	USE
Dep. acl	45	0.0118	0.0147	**0.0117**	0.0125	**0.0117**	0.0119	**0.0117**	0.0118
Dep. compound	82	0.0131	0.016	**0.013**	0.0138	**0.013**	0.0132	**0.013**	**0.013**
Dep. prep	299	0.0115	0.0144	**0.0113**	0.0122	**0.0113**	0.0114	**0.0113**	**0.0113**
Dep. relcl	58	0.0107	0.0137	**0.0106**	0.0114	**0.0106**	0.0108	**0.0106**	**0.0106**
NE ORG	22	0.0105	0.0136	**0.0104**	0.0113	**0.0104**	0.0106	**0.0104**	**0.0104**
POS NN & J & V	228	0.0112	0.0141	**0.011**	0.0119	**0.011**	0.0112	**0.011**	**0.011**
Sentiment (neg.)	24	0.0115	0.0145	0.0115	0.0123	**0.0114**	0.0116	**0.0114**	0.0115
Sentiment (neut.)	310	0.0115	0.0144	0.0114	0.0122	**0.0113**	0.0115	**0.0113**	0.0114
Sentiment (pos.)	45	0.0103	0.0134	0.0103	0.0111	**0.0102**	0.0104	**0.0102**	0.0103

Table 21: Mean of the absolute error averaged across dimensions (MAED), per ling. feature, for the fMRI dataset **Pereira (large)**

	count	BERT	ELMo	fastText	GloVe	InferSent	PM	ST	USE
Dep. acl	14	0.0589	0.06	**0.0588**	0.0589	**0.0588**	0.0589	**0.0588**	**0.0588**
Dep. compound	84	**0.0648**	0.0658	**0.0648**	0.065	**0.0648**	**0.0648**	**0.0648**	**0.0648**
Dep. neg	64	0.0667	0.0675	0.0667	0.0668	**0.0666**	0.0667	**0.0666**	0.0667
Dep. prep	219	0.0643	0.0652	**0.0642**	0.0644	**0.0642**	**0.0642**	**0.0642**	**0.0642**
Dep. relcl	44	**0.0582**	0.0594	**0.0582**	0.0583	**0.0582**	0.0583	**0.0582**	**0.0582**
NE ORG	30	0.0675	0.0684	0.0675	0.0676	**0.0674**	0.0675	**0.0674**	**0.0674**
NE PERSON	197	**0.0679**	0.0688	**0.0679**	0.068	**0.0679**	**0.0679**	**0.0679**	**0.0679**
POS 3+ NN	40	0.0571	0.0582	**0.057**	0.0571	**0.057**	0.0571	**0.057**	**0.057**
POS NN & J & V	139	**0.0634**	0.0644	**0.0634**	0.0635	**0.0634**	**0.0634**	**0.0634**	**0.0634**
Sentiment (neg.)	24	**0.0663**	0.0672	**0.0663**	0.0665	**0.0663**	**0.0663**	**0.0663**	**0.0663**
Sentiment (neut.)	317	0.0727	0.0735	0.0727	0.0728	**0.0726**	0.0727	**0.0726**	**0.0726**
Sentiment (pos.)	35	**0.068**	0.0689	**0.068**	0.0682	**0.068**	**0.068**	**0.068**	**0.068**

Table 22: Mean of the absolute error averaged across dimensions (MAED), per ling. feature, for the fMRI dataset **Wehbe**

	BERT	ELMo	fastText	GloVe 50	InferSent	PM	ST	USE
FFD	0.0697	0.0752	0.0857	0.0881	0.0685	0.084	**0.0681**	0.0726
FP	0.0949	0.1028	0.123	0.1253	**0.0901**	0.112	0.0929	0.1018
MFD	0.0713	0.0771	0.0858	0.0883	0.0699	0.0805	**0.068**	0.0741
nFix	0.0493	0.0506	0.0537	0.0559	0.0453	0.0529	**0.044**	0.0551
TFD/TRT	0.0437	0.0426	0.0467	0.0481	0.0422	0.0457	**0.0397**	0.045

Table 23: Mean AED per feature for Eye-Tracking/Dundee

	BERT	ELMo	fastText	GloVe 50	InferSent	PM	ST	USE
FFD	0.0424	0.0437	0.057	0.0531	0.0425	0.0448	**0.039**	0.0442
GD	0.0285	0.0302	0.0373	0.034	0.0291	0.0306	**0.0279**	0.0406
nFix	0.0305	0.0311	0.0406	0.0362	0.0273	0.0286	**0.0259**	0.0309
TRT	0.0275	0.0288	0.0344	0.0314	0.0265	0.0274	**0.0241**	0.0295

Table 24: Mean AED per feature for Eye-Tracking/GECO

	BERT	ELMo	fastText	GloVe 50	InferSent	PM	ST	USE
FFD	0.0838	0.0912	0.1031	0.107	**0.0782**	0.0967	0.0805	0.0898
GD	0.0622	0.0653	0.071	0.072	0.0582	0.0677	**0.0581**	0.0678
nFix	0.06	0.0607	0.0665	0.0693	**0.0558**	0.0646	0.0607	0.0659
TRT	0.0573	0.061	0.0634	0.0648	**0.0541**	0.0602	0.0544	0.0618

Table 25: Mean AED per feature for Eye-Tracking/ZuCo

	BERT	ELMo	fastText	GloVe	InferSent	PM	ST	USE
M02	**0.0093**	0.0119	0.0099	0.0095	**0.0093**	0.0141	**0.0093**	0.0094
M04	**0.0106**	0.0135	0.0112	0.0173	**0.0106**	0.0153	**0.0106**	0.0108
M07	0.0154	0.0135	0.0114	0.0123	**0.0107**	0.0157	**0.0107**	0.0108
M15	0.0095	0.0123	0.0101	0.0098	0.0095	0.0144	0.0095	**0.0094**
P01	0.0079	0.0095	0.0071	0.0064	**0.0063**	0.012	**0.0063**	**0.0063**

Table 26: Mean AED per subject for fMRI/Pereira (small) (original subject IDs)

	BERT	ELMo	fastText	GloVe	InferSent	PM	ST	USE
M02	0.0079	0.0111	0.0078	0.0088	**0.0077**	0.0079	**0.0077**	**0.0077**
M04	**0.0091**	0.0122	**0.0091**	0.0099	**0.0091**	0.0093	**0.0091**	0.0092
M07	0.0108	0.0134	0.0104	0.0111	**0.0103**	0.0105	**0.0103**	**0.0103**
M08	**0.0149**	0.0177	**0.0149**	0.0159	**0.0149**	0.015	**0.0149**	**0.0149**
M09	0.0123	0.0151	0.0122	0.013	**0.0121**	0.0123	**0.0121**	0.0122
M14	0.0188	0.0212	0.0188	0.0194	**0.0187**	0.0188	**0.0187**	0.0188
M15	0.0103	0.0134	**0.0102**	0.0109	**0.0102**	0.0103	**0.0102**	**0.0102**
P01	0.0068	0.0102	**0.0067**	0.008	**0.0067**	0.0069	**0.0067**	0.0068

Table 27: Mean AED per subject for fMRI/Pereira (large) (original subject IDs)

	BERT	ELMo	fastText	GloVe	InferSent	PM	ST	USE
1	**0.0761**	0.077	**0.0761**	0.0764	**0.0761**	**0.0761**	**0.0761**	**0.0761**
2	**0.075**	0.0758	**0.075**	0.0752	**0.075**	**0.075**	**0.075**	**0.075**
3	0.0703	0.0711	**0.0702**	0.0704	**0.0702**	**0.0702**	**0.0702**	**0.0702**
4	0.0715	0.0723	**0.0714**	0.0716	**0.0714**	**0.0714**	**0.0714**	**0.0714**
5	0.071	0.0717	**0.0709**	0.0711	**0.0709**	0.071	**0.0709**	**0.0709**
6	**0.0667**	0.0675	**0.0667**	0.0668	**0.0667**	**0.0667**	**0.0667**	**0.0667**
7	**0.0758**	0.0766	**0.0758**	0.076	**0.0758**	**0.0758**	**0.0758**	**0.0758**
8	0.0679	0.0686	0.0679	0.068	**0.0678**	0.0679	**0.0678**	**0.0678**

Table 28: Mean AED per subject for fMRI/Wehbe)

	FFD		FP		MFD		TFD		nFix	
	B	**L**	**B**	**L**	**B**	**L**	**B**	**L**	**B**	**L**
BERT	32	912	32	912	32	912	32	912	32	768
ELMo	32	768	32	1536	32	768	16	512	16	2764
fastText (Wiki Sub.)	64	270	32	270	64	270	64	270	32	270
GloVe 50	16	45	16	45	16	45	16	25	16	45
InferSent	32	2048	64	3072	32	3072	32	3072	32	2048
Power-Mean	32	1800	32	1800	32	1800	16	900	32	1800
Skip-Thoughts	32	3600	32	2400	32	2400	32	2400	64	3600
USE (DAN)	16	461	32	461	32	461	16	461	128	7

Table 29: Final parametrization of batch and layer sizes of individual feature experiments of the **Dundee** eye-tracking dataset

	FFD		GD		TRT		nFix	
	B	L	B	L	B	L	B	L
BERT	16	768	16	912	16	912	16	912
ELMo	16	768	32	768	32	768	16	768
fastText (CC Sub.)	16	270	32	270	32	270	32	270
GloVe 50	32	38	32	38	32	38	32	38
InferSent	32	2048	32	2048	32	2048	32	2048
Power-Mean	32	2700	32	2700	32	2700	32	2700
Skip-Thoughts	32	3600	32	3600	32	3600	16	3600
USE (DAN)	16	461	32	461	32	461	32	461

Table 30: Final parametrization of batch and layer sizes of individual feature experiments of the **ZuCo** eye-tracking dataset

	FFD		GD		TRT		nFix	
	B	L	B	L	B	L	B	L
BERT	32	912	32	768	32	912	32	912
ELMo	16	2304	16	2304	16	2765	16	2765
fastText (CC)	128	270	32	270	64	225	64	270
GloVe 50	16	45	16	45	16	45	16	45
InferSent	32	3072	32	3072	32	3072	32	3072
Power-Mean	16	900	16	900	16	450	16	450
Skip-Thoughts	128	2400	64	2400	128	2400	128	2400
USE (DAN)	16	461	128	7	64	461	64	461

Table 31: Final parametrization of batch and layer sizes of individual feature experiments of the **GECO** eye-tracking dataset

	Natural speech		ZuCo		ZuCo2	
	B	L	B	L	B	L
BERT	16	256	32	256	32	256
ELMo	32	384	32	204/307	16	204
fastText (Wiki Sub.)	32	150	32	75	32	30
GloVe 50	32	13	32	13	16	13
InferSent	32	103	32	410	32	103
Power-Mean	32	180	32	180	16	135
Skip-Thoughts	32	180	32	180	32	240
USE (DAN)	32	256	32	256	32	256

Table 32: Final parametrization of batch and layer sizes of EEG experiments

	Pereira (small)		**Pereira (large)**		**Wehbe**	
	Batch	**Layer**	**Batch**	**Layer**	**Batch**	**Layer**
BERT	8	7	16	14	16	14
ELMo	8	768	8	307	8	192
fastText	8	225 (CC)	16	15 (Wiki sub.)	32	225 (CC)
GloVe 50	8	7	8	13	16	38
InferSent	8	7	16	7	32	13
Power-Mean	8	1800	8	6	8	6
Skip-Thoughts	8	7	16	7	32	15
USE (DAN)	16	7	(16)	(14)	(32)	(13)
USE (Transformer)	(16)	(13)	16	13	32	26

Table 33: Final parametrization of batch and layer sizes of fMRI experiments. The best-performing USE model per dataset is chosen (values without parentheses)