# CogniVal: A Framework for Cognitive Word Embedding Evaluation

**Nora Hollenstein[1], Antonio de la Torre[1], Nicolas Langer[2], Ce Zhang[1]**

[1] Department of Computer Science, ETH Zurich

{noraho,antonide,ce.zhang}@ethz.ch

[2] Department of Psychology, University of Zurich

n.langer@psychologie.uzh.ch

## Abstract

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

## 1 Introduction

Word embeddings are the corner stones of state-of-the-art NLP models. Distributional representations which interpret words, phrases, and sentences as high-dimensional vectors in semantic space have become increasingly popular. These vectors are obtained by training language models on large corpora to encode contextual information. Each vector represents the meaning of a word.

Evaluating and comparing the quality of different word embeddings is a well-known, largely open challenge. Currently, word embeddings are evaluated with extrinsic or intrinsic methods. Extrinsic evaluation is the process of assessing the quality of the embeddings based on their performance on downstream NLP tasks, such as question answering or entity recognition. However, embeddings can be trained and fine-tuned for spe-
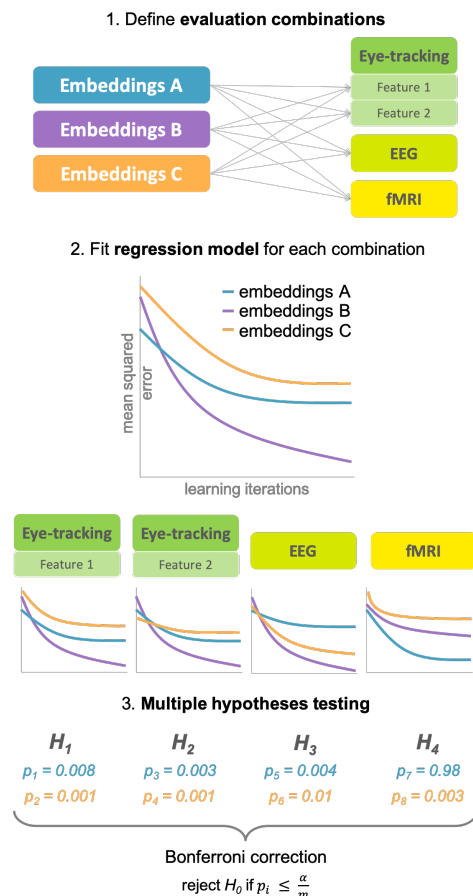


Figure 1: Overview of the cognitive word embedding evaluation process.

cific tasks, but this does not mean that they accurately reflect the meaning of words.

One the other hand, intrinsic evaluation methods, such as word similarity and word analogy tasks, merely test single linguistic aspects. These tasks are based on conscious human judgements. Conscious judgements can be biased by subjective factors and the tasks themselves might also be biased (Malvina Nissim, 2019). Additionally, the correlation between intrinsic and extrinsic metrics is not very clear, as intrinsic evaluation results fail

to predict extrinsic performance (Chiu et al., 2016; Gladkova and Drozd, 2016). Finally, both intrinsic and extrinsic evaluation types often lack statistical significance testing and do not provide a global quality score.

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

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

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

2. **Diversity** within modalities: Evaluate against different datasets within one modality to make sure the number of samples is as large as possible.

3. **Correlation** of results should be evident across modalities and even between datasets of the same modality.

**Contributions** We present CogniVal, the first framework of cognitive word embedding eval-

uation to follow these principles and analyze the findings. We evaluate different embedding types against a combination of 15 cognitive data sources, acquired via three modalities: eye-tracking, electroencephalography (EEG) and functional magnetic resonance imaging (fMRI). The word representations are evaluated by assessing their ability of predicting cognitive language processing data. After fitting a neural regression model for each combination, we apply multiple hypotheses testing to measure the statistical significance of the results, taking into account multiple comparisons (see Figure 1). This contributes to the consistency of the results and to attain a global score of embedding quality. Our main findings when evaluating six state-of-the-art word embeddings with CogniVal show that the majority of embedding types significantly outperform a baseline of random embeddings when predicting a wide range of cognitive features. Additionally, the results show consistent correlations between between datasets of the same modality and across different modalities, validating the intuition of our approach. Finally, we present an exploratory but promising correlation analysis between the scores obtained using our intrinsic evaluation methods and the performance on extrinsic NLP tasks.

The code of this evaluation framework is openly available[1]. It can be used as is, or in combination with other intrinsic as well as extrinsic evaluation methods for word representations.

## 2 Related Work

Mitchell et al. (2008) pioneered the use of word embeddings to predict patterns of neural activation when subjects are exposed to isolated word stimuli. More recently, this dataset and other fMRI resources have been used to evaluate learned word representations.

For instance, Abnar et al. (2018) and Rodrigues et al. (2018) evaluate different embeddings by predicting the neuronal activity from the 60 nouns presented by Mitchell et al. (2008). Søgaard (2016) shows preliminary results in evaluating embeddings against continuous text stimuli in eye-tracking and fMRI data. Moreover, Beinborn et al. (2019) recently presented an extensive set of language–brain encoding experiments. Specifically, they evaluated the ability of an ELMo language model to predict brain responses of multiple

---

[1] https://github.com/DS3Lab/cognival

fMRI datasets.

EEG data has been used for similar purposes. Schwartz and Mitchell (2019) and Ettinger et al. (2016) show that components of event-related potentials can successfully be predicted with neural network models and word embeddings.

However, these approaches mostly focus on one modality of brain activity data from small individual cognitive datasets. The lack of data sources has been one reason why this type of evaluation has not been too popular until now (Bakarov, 2018a). Hence, in this work we collected a wide range of cognitive data sources ranging from eye-tracking to EEG and fMRI to ensure coverage of different features, and consequently of the cognitive processing taking place in the human brain during reading.

**Evidence from cognitive neuroscience** Murphy et al. (2018) review computational approaches to the study of language with neuroimaging data and show how different type of words activate neurons in different brain regions. Similarly, mapping fMRI data from subjects listening to stories to the activated brain regions, revealed semantic maps of how words are distributed across the human cerebral cortex (Huth et al., 2016).

Furthermore, word predictability and semantic similarity show distinct patterns of brain activity during language comprehension: semantic distances can have neurally distinguishable effects during language comprehension (Frank and Willems, 2017). These findings support the theory that brain activity data does reflect lexical semantics and is thus an appropriate foundation for evaluating the quality of word embeddings.

## 3  Word embeddings

Pre-trained word vectors are an essential component in state-of-the-art NLP systems. We chose six commonly used pre-trained embeddings to evaluate against the cognitive data sources. See Table 1 for an overview of the dimensions of each embedding type. We evaluate the following types of word embeddings:

- **Glove**: Pennington et al. (2014) provide embeddings of different dimensions trained on aggregated global word-word co-occurrence statistics over a corpus of 6 billion words.

| embeddings | dim. | hidden layer units |
|---|---|---|
| Glove | 50 | [30, 26, 20, 5] |
| Glove | 100 | [50, 30] |
| Glove | 200 | [100, 50] |
| Glove | 300 | [150, 50] |
| Word2vec | 300 | [150, 50] |
| WordNet2vec | 850 | [400, 200] |
| FastText | 300 | [150, 50] |
| ELMo | 1024 | [600, 200] |
| BERT | 768 | [400, 200] |
| BERT | 1024 | [600, 200] |

Table 1: Overview of word embeddings evaluated with CogniVal. The last column shows the search space of the grid search for the number of units in the hidden layer.

- **Word2vec**: Non-contextual embeddings trained on 100 billion words from a Google News dataset (Mikolov et al., 2013).

- **WordNet2Vec** (Saedi et al., 2018) These embeddings represent the conversion from semantic networks into semantic spaces. Trained on WordNet, a lexical ontology for English that comprises over 155,000 lemmas (but trained only on 60,000 words).

- **FastText** pre-trained embeddings use character n-grams to compose the vector of the full words (Mikolov et al., 2018). We evaluate the embeddings with and without subword information trained on 16 billion tokens of Wikipedia sentences as well as the ones trained on 600 billion tokens of Common Crawl.

- **ELMo** models both complex characteristics of word use (i.e. syntax and semantics), and how these uses vary across linguistic contexts (Peters et al., 2018). These word vectors are learned functions of the internal states of a deep bidirectional language model, which is pre-trained on a large text corpus. We take the first of the three output layers, containing the context insensitive word representations.

- **BERT** embeddings are contextual, bidirectional word representations, based on the idea that fine-tuning a pre-trained language model can help the model achieve better results in the downstream tasks (Devlin et al., 2019). We take the hidden states of the second to last

of 12 output layers as the representation for each token.

# 4 Cognitive data

In this paper, we consider three modalities of recording cognitive language processing signals: eye-tracking, electroencephalography (EEG), and functional magnetic resonance imaging (fMRI). All three methods are complementary in terms of temporal and spatial resolution as well as the directness in the measurement of neural activity (Mulert, 2013). For the word embedding evaluation we selected a wide range of datasets from these three modalities to ensure a more diverse and accurate representation of the brain activity during language processing.

Table 2 shows an overview of the cognitive data sources used, which are described in more detail below. Since the processing in the brain differs depending on whether the information is accessed via the visual or auditory system (Price, 2012), we include data of different stimuli, e.g. participants reading sentences or listening to audio-books. Moreover, our collection of cognitive data sources contains datasets of both isolated (single words) and continuous (words in context, i.e. sentences or stories) stimuli. All datasets include English language stimuli and the participants were native speakers or highly proficient.

**Eye-tracking** Eye-tracking is an indirect measure of cognitive activity. Gaze patterns are highly correlated with the cognitive load associated with different stages of human text processing (Rayner, 1998). For instance, fixation duration is higher for long, infrequent and unfamiliar words (Just and Carpenter, 1980).

All eye-tracking datasets used in this work were recorded from natural, self-paced reading. Each dataset provides different eye-tracking features. The most common features, available in all 7 datasets are: first fixation duration, first pass duration, mean fixation duration, total fixation duration and number of fixations. For a complete list and description of the eye-tracking features available in each corpus see Appendix A.1.

Gaze vectors consist of specific features, which are extracted based on the reading times, fixations and regressions on each word. Feature values are aggregated on word type level and scaled between 0 and 1. The feature values were averaged over all subjects within a dataset. This preprocess-

ing step is done separately for each data source before combining them. Hollenstein and Zhang (2019) show that combining gaze data from different sources can be helpful for NLP applications, even when they are recorded with different devices and filtering,

By using as many features as available from each dataset, including features characterizing basic, early and late word processing aspects, the goal is to cover the whole language understanding process on word level.

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

The EEG datasets used in this work were either recorded from reading sentences or listening to natural speech. Word-level brain activity could be extracted by mapping to eye-tracking cues (ZuCo), by mapping to auditory triggers (NATURAL SPEECH), by recording only the last word in each sentence (N400), or through serial presentation of the words (UCL). Standard preprocessing steps for EEG data, including band-pass filtering and artifact removal, are performed in the same manner for all four data sources. See Appendix A.2 for details on EEG preprocessing.

The EEG data is aggregated over all available subjects and over all occurrences of a token. This yields an $n$-dimensional vector, where $n$ is the number of electrodes, ranging from 32 to 130, depending on the EEG device used to record the data.

EEG data can be aggregated over all subjects within one dataset, because the number and locations of electrodes are identical. However, due to the differences in the number of electrodes between datasets, we cannot aggregate over all EEG datasets.

**fMRI** Functional magnetic resonance imaging is a technique for measuring and mapping brain activity by detecting changes associated with blood flow. fMRI has a temporal resolution of two seconds, which means that with continuous stimuli such as natural reading or story listening, one scan covers multiple words. We use datasets of

| | Data source | stimulus | subj. | tokens | types | coverage |
|---|---|---|---|---|---|---|
| **EYE-TRACKING** | GECO (Cop et al., 2017) | text | 14 | 68606 | 5383 | 95% |
| | DUNDEE (Kennedy et al., 2003) | text | 10 | 58598 | 9131 | 94% |
| | CFILT-SARCASM (Mishra et al., 2016) | text | 5 | 23466 | 4237 | 85% |
| | ZUCO (Hollenstein et al., 2018) | text | 12 | 13717 | 4384 | 90% |
| | CFILT-SCANPATH (Mishra et al., 2017) | text | 5 | 3677 | 1314 | 89% |
| | PROVO (Luke and Christianson, 2017) | text | 84 | 2743 | 1192 | 95% |
| | UCL (Frank et al., 2013) | text | 43 | 1886 | 711 | 98% |
| | ALL EYE-TRACKING (aggregated) | text | - | 26353 | 16419 | 88% |
| **EEG** | ZUCO (Hollenstein et al., 2018) | text | 12 | 13717 | 4384 | 90% |
| | NATURAL SPEECH (Broderick et al., 2018) | speech | 19 | 12000 | 1625 | 98% |
| | UCL (Frank et al., 2015) | text | 24 | 1931 | 711 | 98% |
| | N400 (Broderick et al., 2018) | text | 9 | 150 | 140 | 100% |
| **fMRI** | HARRY POTTER (Wehbe et al., 2014) | text | 8 | 5169 | 1295 | 92% |
| | ALICE (Brennan et al., 2016) | speech | 27 | 2066 | 588 | 99% |
| | PEREIRA (Pereira et al., 2018) | text/image | 15 | 180 | 180 | 99% |
| | NOUNS (Mitchell et al., 2008) | image | 9 | 60 | 60 | 100% |

Table 2: Cognitive data sources used in this work. Coverage is the percentage of vocabulary in data source occurs in British National Corpus list of most frequent English words[2].

isolated stimuli (e.g the NOUNS dataset) as well as continuous stimuli (e.g. HARRY POTTER). While it is easier to extract word-level signals from isolated stimuli, continuous stimuli allow extracting signals in context over a wider vocabulary.

Where multiple trials were available, the brain activation for each word is calculated by taking the mean over the scans. Moreover, if the stimulus is continuous (HARRY POTTER and ALICE datasets), the data is aligned with an offset of four seconds to account for hemodynamic delay[3].

fMRI data contains representations of neural activity of millimeter-sized cubes called voxels. Standard fMRI preprocessing methods such as motion correction, slice timing correction and co-registration had already been applied before. To select the voxels to be predicted we use the pipeline provided by Beinborn et al. (2019). This pipeline consists of extracting corresponding scan(s) for each word, and randomly selecting 100, 500 and 1000 voxels (for the HARRY POTTER, PEREIRA and NOUNS datasets). The published version of the ALICE dataset provided

the preprocessed signal averaged for six regions of interest, hence for this particular dataset we predict the activation for these regions only. Appendix A.3 contains the details of the preprocessing steps. Finally, the fMRI data is converted to $n$-dimensional vectors, where $n$ is the number of randomly selected voxels (100, 500 or 1000) or regions (6).

## 5 Embedding evaluation method

In order to evaluate the word embeddings against human lexical representations, we fit the embeddings to a wide range cognitive features, i.e. eye-tracking features and activation levels of EEG and fMRI. This section describes how these models were trained and evaluated. After evaluating each combination separately, we test for statistical significance taking into account the multiple comparisons problem. See Figure 1 for an overview of the evaluation process.

### 5.1 Models

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

---

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

[3] The fMRI signal measures a brain response to a stimulus with a delay of a few seconds, and it decays slowly over a duration of several seconds (Miezin et al., 2000). For continuous stimuli, this means that the response to previous stimuli will have an influence on the current signal. Thus, context of the previous words is taken into account
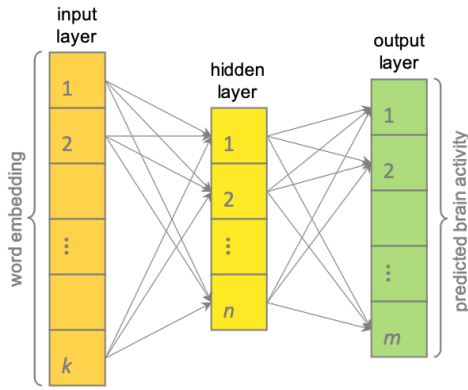
Figure 2: Neural architecture of regression models.

first fixation durations of the DUNDEE corpus.

For the regression models, we train neural networks with $k$ input dimensions, one dense hidden layer of $n$ nodes using ReLU activation and an output layer of $m$ nodes using linear activation. The model is a multiple regression with layers of dimension $k$-$n$-$m$, where $k$ is the number of dimensions of the word embeddings and $m$ changes depending on the cognitive data source to be predicted. For predicting single eye-tracking features $m$ equals 1, whereas for predicting EEG of fMRI vectors $m$ is the dimension of the cognitive data vector, or more specifically, the number of electrodes in the EEG data or the number of voxels in the fMRI data. Figure 2 shows this neural architecture. The loss function optimizes the mean squared error (MSE) and uses an Adam optimizer with a learning rate of 0.001.

5-fold cross validation is performed for each model (80% training data and 20% test data). The optimal number of nodes $n$ in the hidden layer is selected individually for each combination of cognitive data source and embedding type. To this end, a grid search is performed before training, which is evaluated on a validation set consisting of 20% of the training data with 3-fold cross validation (see Table 1 for details on the search space). The best model is then saved and used to predict the cognitive feature for each word in the test set. Finally, the results are measured with the mean squared error, averaged over all predicted words.

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

| | voxels | | |
|---|---|---|---|
| embeddings | 100 | 500 | 1000 |
| glove-300 | 0.119 | 0.081 | 0.078 |
| word2vec | 0.103 | 0.075 | 0.075 |
| fasttext-crawl-sub | 0.092 | 0.070 | 0.069 |
| bert-base | 0.020 | 0.017 | 0.016 |
| wordnet2vec | 0.105 | 0.077 | 0.076 |
| elmo | 0.067 | 0.051 | 0.050 |

Table 3: Effect of predicting different numbers of randomly selected voxels.

| embeddings | nFix | TRT | FFD |
|---|---|---|---|
| glove-300 | 0.010 | 0.017 | 0.027 |
| word2vec | 0.009 | 0.010 | 0.016 |
| fasttext-crawl-sub | 0.008 | 0.007 | 0.012 |
| bert-base | 0.005 | 0.003 | 0.004 |
| wordnet2vec | 0.010 | 0.010 | 0.019 |
| elmo | 0.008 | 0.009 | 0.011 |
| **average** | **0.008** | **0.009** | **0.015** |

Table 4: Comparison of word embeddings predicting single eye-tracking features: number of fixations (nFix), first fixation duration (FFD) and total reading time of a word (TRT).

## 5.2 Multiple hypotheses testing

With the purpose of achieving consistency and going towards a global quality metric that can be combined with other evaluation methods, we perform statistical significance testing on each hypothesis. A hypothesis consists of comparing the combination of an embedding type and a cognitive data source to the random baseline.

Since the distribution of our test data is unknown and the datasets are small, we perform a Wilcoxon signed-rank test for each hypothesis (Dror et al., 2018). Additionally, to counteract the multiple hypotheses problem, we apply the conservative Bonferroni correction, where the global null hypothesis is rejected if $p < \alpha/N$, where $N$ is the number of hypotheses (Dror et al., 2017). In our setting, $\alpha = 0.01$ and $N = 4$ for EEG (one hypothesis per EEG data source), $N = 59$ for for fMRI (one hypothesis per participant of each fMRI data source), and $N = 42$ for eye-tracking (one hypothesis per feature per eye-tracking corpus).

This approach of significance testing can easily be used in combination with other intrinsic and extrinsic evaluation methods. The significance ratios
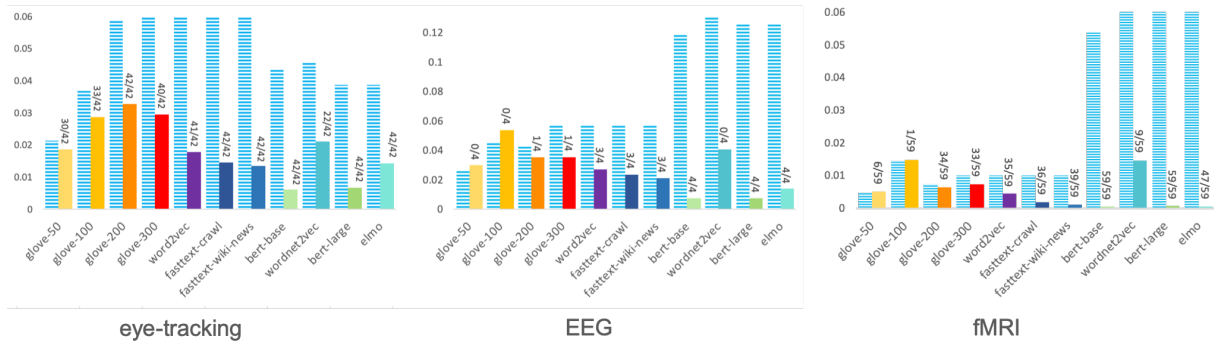
Figure 3: Results for each modality: Aggregated results for all embeddings predicting cognitive features for all datasets of a modality (sorted by dimension of embeddings in increasing order from left to right). The striped blued bars represent random baseline. The labels on the embedding bars show the ration of significant results under the Bonferroni correction to the total number of hypotheses.
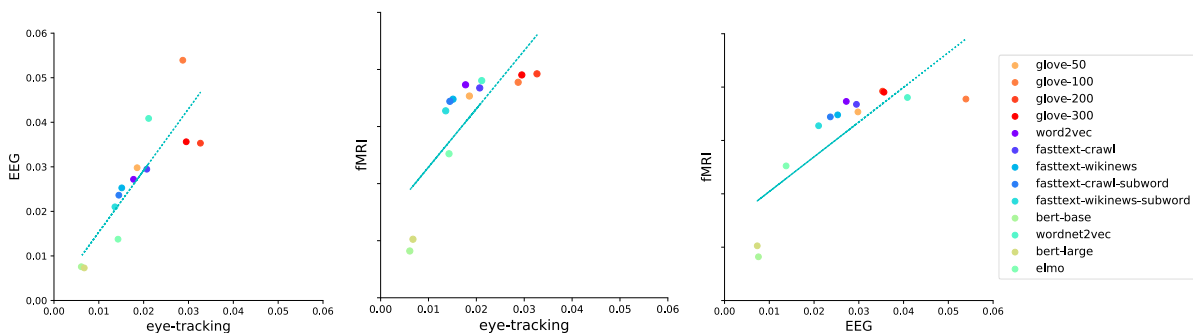


Figure 4: Correlation plots between all three modalities of cognitive signals.

are shown in Figure 3.

## 6 Results & Discussion

**Prediction results** First, we show in Figure 3 how well each word embedding type is able to predict eye-tracking features, EEG and fMRI vectors. As can be seen the majority of results are significantly better than the random baselines. BERT, ELMo and FastText embeddings achieve the best prediction results. All exact numbers can be found in Appendix B. While a random baseline can be considered a rather naive choice, this setting also allows us compare the performance between word embedding types.

When predicting single eye-tracking features, the performance varies greatly. For instance, Table 4 shows that the prediction error on number of fixations and total reading time from the ZuCo dataset is much lower than for first fixation duration. This suggest that more general eye-tracking features covering the complete reading process of a word are easier to predict.

In the case of predicting voxel vectors of fMRI data, the results improve when choosing a larger number of voxels (see Table 3). Hence, in the remainder of this work we present only the results for 1000 voxels.

We also examined the EEG results in more depth by analyzing which electrodes are predicted more accurately and which electrodes values are very difficult to predict. This is exemplified by Figure 5, which shows the 20 best and worst predicted electrodes of the ZuCo data for the BERT embeddings of 1024 dimensions as well as aggregated over all cognitive data sources. The middle central electrodes are predicted more accurately. The middle central electrodes are known to register the activity of the Perisylvian cortex, which is relevant for language related processing (Catani et al., 2005). Moreover it can be speculated that there is a frontal asymmetry between the electrodes on the left and right hemispheres.

**Cognitive data implications** The diversity of cognitive data sources chosen for this work allows us to analyze and compare results on several levels and between several cognitive metrics. In order to conduct this evaluation on a collection of 15
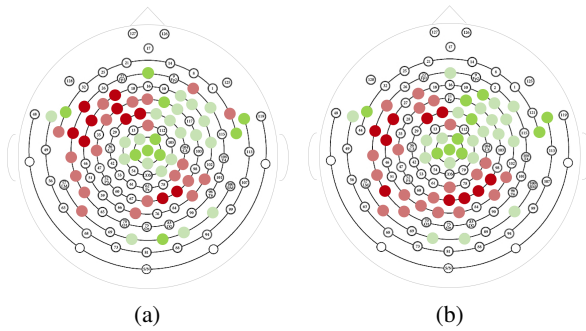
Figure 5: EEG electrode analysis, (a) for BERT (large) and (b) aggregated over all embedding types. Red = worst predicted electrodes, green = best predicted electrodes.



Figure 6: Correlation between results on EEG datasets.

datasets from three modalities, many crucial decisions were taken about preprocessing, feature extraction and evaluation type. Since there are different methods on how to process different types of cognitive language understanding signal, it is important to make these decisions transparent and reproducible.

Moreover, it is a challenge to segment brain activity data correctly and meaningfully into word-level signal from naturalistic, continuous language stimulus (Hamilton and Huth, 2018). This makes consistent preprocessing across data sources even more important.

Another challenge is to consolidate the cognitive features to be predicted. For instance, we chose a wide selection of eye-tracking features that cover early and late word processing. However, choosing only general eye-tracking features such as total reading time would also be a viable strategy. On the other hand, the EEG evaluation could be more coarse-grained, one could also try to predict known ERP effects (e.g. Ettinger et al. (2016)) or features selected based on frequency bands. Moreover, the voxel selection in the fMRI preprocessing could be improved by either predicting all voxels or applying information-driven voxel selection methods (Beinborn et al., 2019).

**Correlations between modalities** Next, we analyze the correlation between the predictions of the three modalities (Figure 4). There is a strong correlation between the results of predicting eye-tracking, EEG and fMRI features. This implies that word embeddings are actually predicting brain activity signals and not merely preprocessing artifacts of each modality. Moreover, the
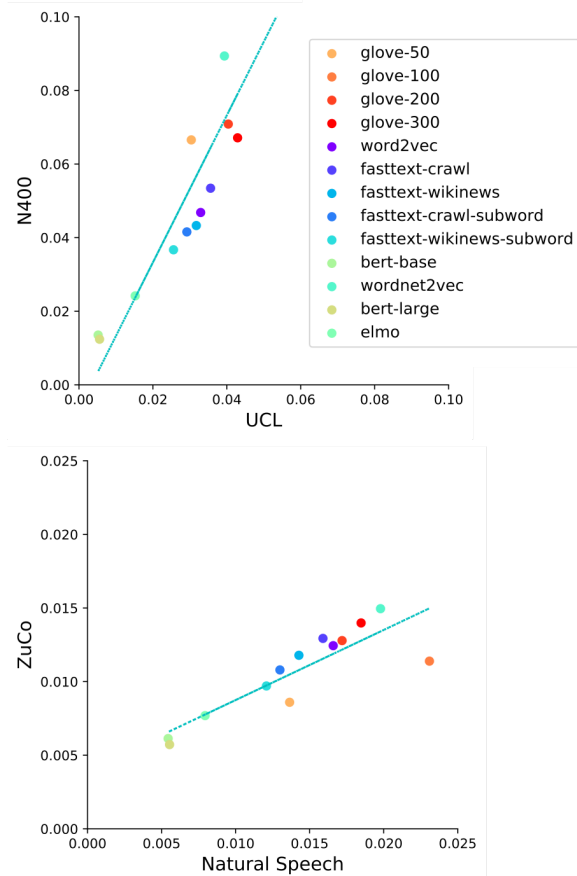
same correlation is also evident between individual datasets within the same modality. As an example, Figure 6 (bottom) shows the correlation of the results predicted for the Natural Speech and ZuCo EEG datasets, where the first had speech stimuli and the latter text. Figure 6 (top) reveals the same positive correlation for two EEG datasets that were preprocessed differently and were recorded with a different number of electrodes. Moreover, the UCL dataset contains word-by-word reading and the N400 contains natural reading of full sentences.

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

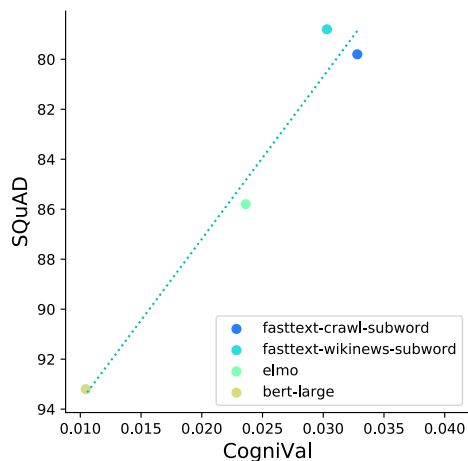The SQuAD results are taken from Devlin et al.

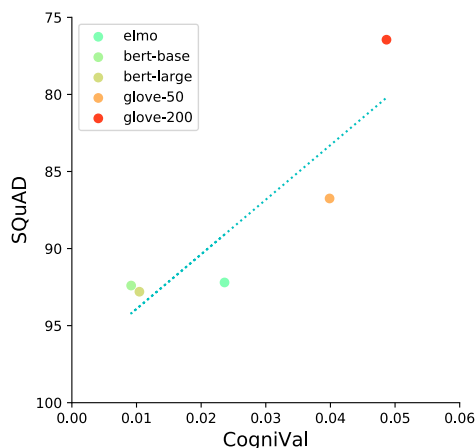Figure 7: Correlation between the SQuAD 1.1 task and the CogniVal results.



Figure 8: Correlation between NER on CoNLL-2003 and the CogniVal results.

(2019) for BERT, from Mikolov et al. (2018) for FastText, and from Peters et al. (2018) for ELMo. The NER results are from the same source for ELMo and BERT, for Glove-50 from Pennington et al. (2014) and for Glove-200 from Ghannay et al. (2016). We correlated these results to the prediction results over all cognitive data sources. Figures 7 and 8 show the correlation plots between the CogniVal results and the two downstream tasks.

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

# 7   Conclusion

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

We plan to expand the collection of cognitive data sources as more of them become available, including data from other languages such as the Narrative Brain Dataset (Dutch, fMRI, Lopopolo et al. (2018)) or the Russian Sentence Corpus (eye-tracking, Laurinavichyute et al. (2017)). Thanks to naturalistic recording of longer text spans, Cogni-Val can also be extended to evaluate sentence embeddings or even paragraph embeddings.

CogniVal can become even more effective by combining the results with other intrinsic or extrinsic embedding evaluation frameworks (Nayak et al., 2016; Rogers et al., 2018) and building on the multiple hypotheses testing.

# 8   Acknowledgements

# References

Samira Abnar, Rasyan Ahmed, Max Mijnheer, and Willem Zuidema. 2018. Experiential, distributional and dependency-based word embeddings have complementary roles in decoding brain activity. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 57–66.

Amir Bakarov. 2018a. Can eye movement data be used as ground truth for word embeddings evaluation? *arXiv preprint arXiv:1804.08749*.

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

Lisa Beinborn, Samira Abnar, and Rochelle Choenni. 2019. Robust evaluation of language-brain encoding experiments. *arXiv preprint arXiv:1904.02547*.

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

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

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

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

Uschi Cop, Nicolas Dirix, Denis Drieghe, and Wouter Duyck. 2017. Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior research methods*, 49(2):602–615.

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

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

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

Allyson Ettinger, Naomi Feldman, Philip Resnik, and Colin Phillips. 2016. Modeling N400 amplitude using vector space models of word representation. In *CogSci*.

Stefan L Frank. 2017. Word embedding distance does not predict word reading time.

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

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

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

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

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

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

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

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

Nora Hollenstein and Ce Zhang. 2019. Entity recognition at first sight: Improving NER with eye movement information. In *NAACL*.

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

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

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

AK Laurinavichyute, Irina A Sekerina, SV Alexeeva, and KA Bagdasaryan. 2017. Russian Sentence Corpus: Benchmark measures of eye movements in reading in Cyrillic.

Alessandro Lopopolo, Stefan L Frank, Antal Van den Bosch, Annabel Nijhof, and Roel M Willems. 2018. The Narrative Brain Dataset (NBD), an fMRI dataset for the study of natural language processing in the brain. In *LREC 2018 Workshop on Linguistic and Neuro-Cognitive Resources (LiNCR)*. LREC.

Steven G Luke and Kiel Christianson. 2017. The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, pages 1–8.

Rob van der Goot Malvina Nissim, Rik van Noord. 2019. Fair is better than sensational: Man is to doctor as woman is to doctor. *arXiv preprint arXiv:1905.09866*.

Francis M Miezin, L Maccotta, JM Ollinger, SE Petersen, and RL Buckner. 2000. Characterizing the hemodynamic response: effects of presentation rate, sampling procedure, and the possibility of ordering brain activity based on relative timing. *Neuroimage*, 11(6):735–759.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

George A Miller and Christiane Fellbaum. 1992. Wordnet and the organization of lexical memory. In *Intelligent tutoring systems for foreign language learning*, pages 89–102. Springer.

Abhijit Mishra, Diptesh Kanojia, and Pushpak Bhattacharyya. 2016. Predicting readers' sarcasm understandability by modeling gaze behavior. In *AAAI*, pages 3747–3753.

Abhijit Mishra, Diptesh Kanojia, Seema Nagar, Kuntal Dey, and Pushpak Bhattacharyya. 2017. Scanpath complexity: Modeling reading effort using gaze information. In *AAAI*, pages 4429–4436.

Tom M Mitchell, Svetlana V Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L Malave, Robert A Mason, and Marcel Adam Just. 2008. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195.

Christoph Mulert. 2013. Simultaneous EEG and fMRI: towards the characterization of structure and dynamics of brain networks. *Dialogues in clinical neuroscience*, 15(3):381.

Brian Murphy, Leila Wehbe, and Alona Fyshe. 2018. Decoding language from the brain. *Language, cognition, and computational models*, page 53.

Neha Nayak, Gabor Angeli, and Christopher D Manning. 2016. Evaluating word embeddings using a representative suite of practical tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 19–23.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. 2018. Toward a universal decoder of linguistic meaning from brain activation. *Nature communications*, 9(1):963.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL*.

Cathy J Price. 2012. A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. *Neuroimage*, 62(2):816–847.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

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

Joao António Rodrigues, Ruben Branco, João Silva, Chakaveh Saedi, and António Branco. 2018. Predicting brain activation with WordNet embeddings. In *Proceedings of the Eight Workshop on Cognitive Aspects of Computational Language Learning and Processing*, pages 1–5.

Anna Rogers, Shashwath Hosur Ananthakrishna, and Anna Rumshisky. 2018. What's in your embedding, and how it predicts task performance. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2690–2703.

Chakaveh Saedi, António Branco, João António Rodrigues, and João Silva. 2018. WordNet embeddings. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 122–131.

Dan Schwartz and Tom Mitchell. 2019. Understanding language-elicited EEG data by predicting it from a fine-tuned language model. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 43–57.

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

Erik F Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the 7th Conference on Natural Language Learning*, volume 4, pages 142–147.

Leila Wehbe, Ashish Vaswani, Kevin Knight, and Tom Mitchell. 2014. Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 233–243.