Workshop on Linguistic and Neurocognitive Resources (LiNCr2020), pages 15–27
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Towards best practices for leveraging human language processing signals for natural language processing

Nora Hollenstein[1], Maria Barrett[2], Lisa Beinborn[3]
[1] ETH Zurich, noraho@ethz.ch
[2] University of Copenhagen, mjb@di.ku.dk
[3] University of Amsterdam, l.m.beinborn@uva.nl

Abstract

NLP models are imperfect and lack intricate capabilities that humans access automatically when processing speech or reading a text. Human language processing data can be leveraged to increase the performance of models and to pursue explanatory research for a better understanding of the differences between human and machine language processing. We review recent studies leveraging different types of cognitive processing signals, namely eye-tracking, M/EEG and fMRI data recorded during language understanding. We discuss the role of cognitive data for machine learning-based NLP methods and identify fundamental challenges for processing pipelines. Finally, we propose practical strategies for using these types of cognitive signals to enhance NLP models.

Keywords: cognitive NLP, neurolinguistic resources, eye-tracking, EEG, MEG, fMRI

1. Introduction

Machine learning methods for natural language processing (NLP) are imperfect and still lack the intricate capabilities that humans access automatically when processing speech or reading a text. For instance, humans are able to resolve coreferences and to perform natural language inference, while machine learning methods are not nearly as good (Wang et al., 2019). Human language processing data can be recorded and used to increase the performance of NLP models and to pursue explanatory research in understanding which "human-like" skills our models are still missing.

Linking brain activity and machine learning can increase our understanding of the contents of brain representations, and consequently in how to use these representations to understand, improve and evaluate machine learning methods for NLP. Our aim in this paper is to find common patterns and approaches that have been implemented successfully when leveraging human language processing signals for NLP. The main objective is to guide researchers when navigating the challenges that are unavoidable when working with cognitive data sources.

In recent years, an increasing number of studies using human language processing for improving and evaluating NLP models have emerged. However, consistent practices in pre-processing, feature extraction, and using the human data in the models have not yet been established. Physiological and neuroimaging data is inherently noisy and may also be subject to idiosyncrasy, which makes it more difficult to effectively apply machine learning algorithms. For example, in eye-tracking, an extended fixation duration indicates more complex cognitive processing, but it is not obvious *which* process is occurring. Brain imaging signals help to better locate cognitive processes in the brain, but it is difficult to disentangle the signal pertinent to the task of interest from the noise related to other cognitive processes which are irrelevant for language processing (e.g., motor control, vision, etc.).

In this paper, we review recent NLP studies leveraging different types of human language processing signals, namely eye-tracking, electroencephalography (EEG), magnetoencephalography (MEG), and functional magnetic resonance imaging (fMRI) recorded during language understanding. We discuss the role of cognitive data for machine learning-based NLP methods and identify fundamental challenges for processing pipelines. Based on this discussion, we propose practical strategies for using these types of cognitive signals to augment NLP models. Finally, we explore the ethical considerations of working with human data in NLP.

2. Cognitive signals

In this section, we introduce eye-tracking, EEG, MEG and fMRI as recording techniques of cognitive signals. We describe the technical details and methodological challenges for each technique and discuss how the signals have been used to improve NLP models.

2.1. Eye-tracking

Eye-tracking signals are recorded with a device that tracks the eye movements in a non-intrusive way, most commonly using infra-red light and a camera. Depending on the sampling rate of the recording device, it provides very fine-grained temporal records of one or both eyes.

When a skilled reader reads, the eyes move rapidly from one word to the next, sequentially fixating through the text. Some words are not fixated at all due to an intricate interplay of preview and predictability effects, and some words are fixated several times due to factors such as syntactic reanalysis. The fact that some words are fixated several times makes it possible to study several stages of linguistic cognitive processing. Early gaze measures capture lexical access and early syntactic processing and are based on the first time a word is fixated. Late measures reflect the late syntactic (re-)processing and general disambiguation. These features occur in words that are fixated more than once. Around 10–15% of the fixations are regressions, where the eye focus jumps back to re-read a part of the text.

NLP task	Earliest reference
Part-of-speech tagging	Barrett et al. (2016a)
Sentiment analysis	Mishra et al. (2017b)
Named entity recognition	Hollenstein & Zhang (2019)
Relation detection	Hollenstein et al. (2019a)
Sarcasm detection	Mishra et al. (2016)
Multiword expressions	Rohanian et al. (2017)
Referential/non-referential *it*	Yaneva et al. (2018)
Coreference resolution	Cheri et al. (2016)
Sentence compression	Klerke et al. (2016)
Predicting misreadings	Bingel et al. (2018)
Predicting native language	Berzak et al. (2017)
Predicting language proficiency	Kunze et al. (2013)
Dependency parsing	Strzyz et al. (2019)
Text summarization	Xu et al. (2009)

Table 1: Overview of NLP tasks where eye movements showed improvements along with the earliest reference.

Each fixation lasts on average around 200 ms, but the variation is large and the duration of each fixation has shown to be reliably linked to many word attributes: syntactic, semantic, and discourse-related. The fixation duration can thus be taken as a proxy for cognitive processing. It is out of the scope of this paper to dig into experimental findings, but Rayner (1998) provides an extensive survey. This psycholinguistic line of research has established a range of eye movement features enabling the study of both early and late cognitive textual processing.

Eye-tracking signals in NLP

Eye movement data has successfully been leveraged to improve a wide range of NLP tasks on several text levels, from part-of-speech tagging (Barrett et al., 2016a) to text summarization (Xu et al., 2009). Table 1 shows an overview of the earliest references for each NLP task.

In NLP, the eye tracking signal can be incorporated into models by using the scanpath which denotes the entire fixation trajectory over a text span. Scanpaths can reveal syntactic re-analysis, text difficulty, and other comprehension problems. Larger-scale computational approaches include Klerke et al. (2018), Von der Malsburg and Vasishth (2011), Wallot et al. (2015). Furthermore, Mishra et al. (2017a) learned the gaze representation in a convolutional neural network directly from the scanpath instead of manually selecting features. This might be a promising approach to increase the amount of gaze data available for training and avoid feature engineering.

Challenges in recording eye tracking signals

While low-cost eye-trackers and webcam-based software (e.g., Papoutsaki et al. (2016)) have recently entered the market, performance evaluations have shown that low cost models have a much higher data loss (Funke et al., 2016). Dalmaijer (2014) and Gibaldi et al. (2017) find accuracy and precision acceptable but they mention the low sampling rate as a constraint for research. Reading research using eye movements are dependent on high sampling rate and good – not just *acceptable* – accuracy and precision. While lower precision can be compensated for with larger font sizes and using only the central part of the screen, it does not seem like the current low-cost models are recommendable for reading research due to these factors. Especially when building a large corpus it is worth considering that any validity or reliability loss such as systematic bias (for example, degrading in precision and accuracy towards the periphery of the screen), as well as unsystematic bias (low data quality due to low sampling rate or large data loss), will propagate to all works using this resource.

2.2. EEG & MEG

The electrical activity of neurons in the brain produces currents spreading through the head. These currents also reach the scalp surface, and the resulting voltage fluctuations on the scalp can be recorded as the electroencephalogram (EEG). The neuronal currents inside the head produce magnetic fields which can be measured above the scalp surface as the magnetoencephalogram (MEG). EEG signals reflect electrical brain activity with millisecond-accurate temporal resolution, but poor spatial resolution. Magnetic fields are less distorted than electric fields by the skull and scalp, which results in a better spatial resolution for MEG.

EEG & MEG signals in NLP

EEG signals have achieved fairly good results for classifying mental tasks (e.g., Zhang et al. (2018)) or text difficulty (Chen et al., 2012). Moreover, Parthasarathy and Busso (2017) presented a multi-task learning architecture for classifying emotions from auditory EEG stimulus. Additionally, Murphy and Poesio (2010) detect semantic categories (i.e. types of nouns, binary classification) from simultaneous EEG and MEG recordings, and found MEG to be more informative for this specific task.

However, there is not much work in higher-level semantic or syntactic NLP tasks with larger number of classes due to the low signal-to-noise ratio. Hollenstein et al. (2019a) achieved only modest improvements when using EEG data for sentiment analysis, relation extraction and named entity recognition. For a review on the use of EEG signals for different classification tasks, including an overview of the ML methods, the artifact pre-processing strategies, and the input features, see Craik et al. (2019).

Further, there has been some work in understanding the parallels between machine and EEG language processing signals. For instance, Hale et al. (2018) showed that neural grammar models are able to learn some of the language processing effects that are manifested in EEG. Moreover, Wehbe et al. (2014b) were the first to align word-by-word MEG activity with embeddings from a recurrent neural language model. Schwartz et al. (2019) use MEG and fMRI to fine-tune a BERT language model (Devlin et al., 2019) and showed that the relationship between language and brain activity learned by BERT during this fine-tuning, transfers across multiple participants and performs well on downstream NLP tasks. In a similar fashion, Toneva and Wehbe (2019) compare and interpret word and sequence embeddings from various recent language models on word-by-word MEG and fMRI recordings.

Challenges in processing EEG & MEG signals

MEG and EEG data contain a large ratio of noise as well as signals from other non-language-related processes, but syntactic and semantic text processing is also known to contribute to the signal. Since EEG merely records signals on the brain surface, it is difficult to draw conclusions about which brain regions are more or less helpful for NLP models. MEG allows to localize the magnetic fields to their sources within the brain with good spatial resolution.

The main challenge lies in cleaning the M/EEG recordings and extracting only the signals containing language processing information. First, artifacts from motor and ocular activities have to be removed. Recently, these tedious manual inspection and cleaning steps have been automatized (e.g., Pedroni et al. (2019)), and efforts to unfold the electrophysiological responses from overlapping, continuous stimuli are being introduced (Ehinger and Dimigen, 2019).

Neuroscientists have studied in detail how to filter the M/EEG data based on certain effects occurring during language understanding, and the activity occurring in certain frequency bands. Two popular ways to analyze the EEG signal are power spectrum analysis and event-related potentials (ERPs).

In power spectrum analyses, the average power of a signal in a specific frequency range is computed. The EEG signal is decomposed into functionally distinct frequency bands. These frequency ranges, which are fixed ranges of wave frequencies and amplitudes over a time scale, are known to correlate with certain cognitive functions. Theta activity (4–8 Hz) reflects cognitive control and working memory (Williams et al., 2019); alpha activity (8–12 Hz) has been related to attentiveness (Klimesch, 2012); beta frequencies (12–30 Hz) affect decisions regarding relevance, for instance, in term relevance tasks for information retrieval (Eugster et al., 2014): and gamma-band activity (30–100 Hz) has been used to detect emotions (Li and Lu, 2009). Hypotheses about the role of the various M/EEG frequency bands in language processing and more general cognitive function are a first step, but more work is needed to establish stronger hypotheses linking language to specific frequencies (Alday, 2019).

Secondly, ERPs are measured brain responses that are the direct result of a specific sensory, cognitive, or motor event. For instance, the N400 component, which peaks $\sim$400ms after the onset of the stimulus, is part of the normal brain response to words and other meaningful stimuli (Kutas and Federmeier, 2000). Brouwer et al. (2017) presented a neuro-computational model based on recurrent neural networks, that successfully simulates the N400 and P600 amplitude in language comprehension. To the best of our knowledge, it has not yet been studied how useful ERP features are for improving natural language understanding tasks.

2.3. FMRI

FMRI is a neuroimaging technique that measures brain activity by the changes in the oxygen level of the blood. This technique relies on the fact that cerebral blood flow and neuronal activation are coupled: When a brain area is in use, blood flow to that area increases.

FMRI produces 3D scans of the brain with high spatial resolution of the signal. For statistical analyses, the brain scan is fragmented into voxels which are cubes of constant size. The signal is interpreted as an activation value for every voxel. The number of voxels varies depending on the precision of the scanner and the size and shape of the participant's brain. The voxel location can be identified with 3-dimensional coordinates, but the signal is commonly processed as a flattened vector which ignores the spatial relationships between the voxels. This rather naive modeling assumption simplifies the signal, but might lead to cognitively and biologically implausible findings.

Most publicly available fMRI datasets have already undergone common statistical filters. These pre-processing steps correct for motion of the participant's head, account for different timing of the scan slices and adjust linear trends in the signal (Wikibooks, 2020). In addition, the scans of the individual brains (which vary in size and shape) need to be aligned with a standardized template to group voxels into brain regions and allow for comparisons across subjects. Researchers using datasets that have been collected and published by another lab should be aware of the effect of these probabilistic corrections. They are necessary to further analyze the signal, but might also systematically add noise to the data and lead to misinterpretations.

FMRI signals in NLP

In their pioneering work, Mitchell et al. (2008) measure the brain signal of nine human participants who are instructed to think about a concept. They average the signal for each of the 60 concepts over multiple trials. Their analysis results indicate that it is possible to distinguish between the correct and a random scan by computationally modeling the relations between concepts. Their dataset has become an evaluation benchmark to compare the cognitive plausibility of different word representation models (Fyshe et al., 2014; Søgaard, 2016; Abnar et al., 2018; Anderson et al., 2017; Bulat et al., 2017). The presentation of individual concepts has the advantage that the signal can be directly linked to the experimental stimulus, but the experimental setup is very artificial compared to authentic language processing scenarios. Recently, fMRI datasets involving more naturalistic language stimuli such as sentences (Pereira et al., 2018) and even full stories (Wehbe et al., 2014a; Brennan, 2016; Huth et al., 2016; Dehghani et al., 2017) have been recorded and facilitate contextualized modeling of language processing.[1]

Besides using fMRI signals to better understand and evaluate the structure of computational models of language, the signal has also been used to directly improve the performance on NLP tasks. Bingel et al. (2016) enrich a model for PoS induction with fMRI signals, Li et al. (2018) perform pronoun resolution, and Vodrahalli et al. (2018) classify movie scene annotations. Recently, Toneva and Wehbe (2019) showed that when the language model BERT (Devlin et al., 2019) is fine-tuned to align with brain recordings, it performs better at syntactic tasks such as

[1] Not all of these datasets are publicly available.

subject-verb agreements. These result indicate a transfer of knowledge from human language processing to NLP tasks. So far, the reported improvements are very small and have not yet been verified on other datasets.

Challenges in processing fMRI signals

As it takes several seconds to complete a full scan of the brain, the measured brain response cannot provide high temporal resolution. In addition, the hemodynamic response to a stimulus can only be measured with a delay of several seconds (Miezin et al., 2000) and it decays slowly. As a consequence, it is not possible to directly align fMRI responses with single words when they are presented as continuous stimuli. The delay can be modeled using hemodynamic response functions or more complex modeling techniques, but they do not work equally well in all areas of the brain (Shain et al., 2019). It has not yet been investigated conclusively whether the fMRI signal is temporally fine-grained enough to detect syntax processing signals in the human brain. Gauthier and Levy (2019) showed experiments where only local grammatical dependencies can be decoded. However, Brennan et al. (2016) showed that for the right features (i.e. a count of tree nodes in a probabilistic context-free grammar model) fMRI is fast enough. More recent NLP studies avoid word-level alignment of fMRI data and analyse longer sequences of words instead (Schwartz et al., 2019; Abnar et al., 2019).

The large number of voxels in the fMRI representation leads to a very high-dimensional signal, but the number of stimuli is usually very small for machine learning standards. In order to fit a model, the dimensionality of the signal needs to be reduced because analysis methods such as correlation or similarity metrics often lead to unintuitive results when applied in high-dimensional spaces (Aggarwal et al., 2001). From a processing perspective, data-driven dimensionality reduction methods on the training set are most attractive because they can work on the raw signal and do not rely on theory-driven assumptions (Kriegeskorte et al., 2006). Examples are classification metrics such as explained variance which capture how much information a voxel contributes to a specific task (as in LaConte et al. (2003) and Michel et al. (2011)). Another option are dimensionality reduction methods such as principal component analysis which reduce the dimensions while retaining most of the variance between responses (Gauthier and Levy, 2019).

Unfortunately, existing fMRI datasets for language processing are not yet large enough to enable direct representation learning, for example using autoencoders (Huang et al., 2017; Rowtula et al., 2018). Instead, the signal is often restricted to voxels that fall within a pre-selected set of regions. These regions are commonly selected in a theory-driven manner based on neurolinguistic studies (Brennan et al., 2016; Wehbe et al., 2014a). Fedorenko et al. (2010) proposed a method to selects regions of interest functionally, i.e. pooling of data from corresponding functional regions across subjects. For instance, Abnar et al. (2019) only include the voxels from the top k regions that are most similar across different subjects given the same stimuli.

Due to the technical requirements, fMRI studies mostly use only a small set of stimuli which makes it hard to evaluate the effect size and the generalizability of the results (Hamilton and Huth, 2018). Minnema and Herbelot (2019) perform experiments with additional data, which also lead to the conclusion that there is simply not enough training data available yet to learn a precise mapping. Furthermore, experimental results are commonly not validated on additional datasets to ensure a more robust evaluation (Beinborn et al., 2019).

3. General challenges

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

3.1. Choosing the type of cognitive signals

An important aspect to take into account when choosing a type of cognitive signal is the linguistic level on which the signals are required: from word level, over phrase and sentence level to discourse level. Due to the low temporal resolution and the hemodynamic lag of fMRI, it is more appropriate to use eye-tracking or EEG of MEG data to extract word-level signals in continuous stimuli. Moreover, if using multiple datasets from the same recording modality, it is crucial to ensure proper pre-processing has been conducted on the datasets, or to apply the same pre-processing steps to all datasets.

Eye-tracking, as an indirect metric of cognitive load during the different stages of reading processing, has numerous advantages. It is an accessible method to record millisecond-accurate eye movements and has successfully been leveraged to improve a wide range of NLP tasks on different text processing levels (see Table 1 for an overview). While the improvements on precision and recall are modest, they are consistent across tasks. The impressive body of psycholinguistic research, a range of established metrics, and the intuitive linking from features to words speak in favour of using eye-tracking for NLP.

EEG is another recording technique with very high temporal resolution (i.e. resulting in multiple samples per second). However, as the electrodes measure electrical activity at the surface of the brain – through the bone – it is difficult to know exactly in which brain region the signal originated. EEG signals have been used frequently for classification in brain-computer-interfaces (e.g., classifying text difficulty for speech recognition (Chen et al., 2012)), but have rarely been used to improve NLP tasks (Hollenstein et al., 2019a). Moreover, there are still many open questions regarding which EEG features are most appropriate, and not much EEG data from naturalistic reading is yet openly

available. MEG, however, yields better temporal *and* spatial resolution, which makes is very suitable for NLP. Unfortunately not many MEG datasets from naturalistic studies are currently available.

Finally, the fMRI signal exhibits opposite characteristics. Due to the precise 3D scans, the spatial resolution is very high; but, since it takes a few seconds to produce a scan over the full brain, the temporal resolution is very low. Recently, fMRI data has become popular in NLP to evaluate neural language models (e.g., Schwartz et al. (2019)) and to improve word representations (Toneva and Wehbe, 2019). It is useful to leverage fMRI signals if the localization of cognitive processes plays an important role and to investigate theories about specialized processing areas. Unfortunately fMRI scans are less accessible and more expensive.

Evidently, human language processing recordings are very noisy. Therefore, if possible it is advisable to work with multiple datasets of the same modality, or to work with multiple modalities to achieve more robust results.

It is insightful to run experiments on multiple cognitive datasets of the same modality. This ensures that the NLP models are not merely picking up on the noise in the cognitive data, but actually learning from language processing specific signals. For instance, Hollenstein and Zhang (2019) combine gaze feature from three corpora, and Mensch et al. (2017) learn a shared representation across many fMRI datasets.

Working with data from multiple modalities is also recommendable. For instance, Schwartz et al. (2019) used both MEG and fMRI data to inform language representations, and were able to show how using both modalities simultaneously improves their predictions. Furthermore, Hollenstein et al. (2019b) presented a framework for cognitive word embedding evaluation, where embeddings are evaluated by predicting eye-tracking, EEG and fMRI signals from 15 different datasets. Their results show clear correlations between these three modalities. Barrett et al. (2018b) combined eye-tracking features with prosodic features, keystroke logs from different corpora, and pre-trained word embeddings for part-of-speech induction and chunking. Several methods were used to project the features into a shared feature space and canonical correlation analysis yielded the best results (Faruqui and Dyer, 2014).

Some studies provide data from multiple modalities recorded at different times on different subjects, but on the same stimulus: For example, the UCL corpus (Frank et al., 2013) contains self-paced reading times and eye-tracking data, and was later extended with EEG data (Frank et al., 2015). Similarly, self-paced reading times and fMRI were recorded for the Natural Stories Corpus (Futrell et al., 2018; Shain et al., 2019); EEG and fMRI were recorded for the Alice corpus (Brennan et al., 2016; Hale et al., 2018).

For some sources, data from co-registration studies is available, which means two modalities were recorded simultaneously during the same experiment. This has become more popular, since all three modalities are complementary in terms of temporal and spatial resolution as well as the directness in the measurement of neural activity (Mulert, 2013). Recent reports attest to the feasibility of co-registration studies for studying the neurobiology of natural reading (see Kandylaki and Bornkessel-Schlesewsky (2019) for a review). For example, eye-tracking and EEG recorded concurrently during reading (Dimigen et al., 2011; Henderson et al., 2013; Hollenstein et al., 2018; Hollenstein et al., 2019c) and concurrent eye-tracking and fMRI (Henderson et al., 2015; Henderson et al., 2016). Using data from co-registration studies in NLP allows for comparison on the same language stimuli, on the same population, and on the same language understanding task, where only the recording method differs.

Finally, the presented recording modalities of cognitive signals in this paper are complementary to each other, the information provided by each modality adds to the full picture. Hence, whether co-registration studies are leveraged or simply data from multiple sources and multiple modalities, it is highly recommendable to test all experiments to improve NLP models on more than one dataset and/or modality.

3.2. Selecting a dataset

Datasets of human language processing signals should be chosen based on the research question. It is important to decide whether controlled experiments with clearly distinguishable conditions are required, for instance, if infrequent linguistic phenomena are of interest, or if natural stimuli are favorable to analyze real-world language (Hamilton and Huth, 2018).

As a example for controlled settings, Mitchell et al. (2008) recorded fMRI data from a isolated word stimuli of 60 concrete nouns. In reading studies, serial presentation of words has often been applied, where one word is presented at the time on the screen (e.g., Wehbe et al. (2014a), Frank et al. (2015)). In an EEG dataset provided by Broderick et al. (2018), the participants also read sentences presented word-by-word. Half of the sentences ended with a congruent word and the other half with an incongruent word, so that the difference in the N400 components could be analyzed. This manipulation facilitates the processing and isolation of the cognitive signals, but it does not reflect processes of natural reading, in which the reader has access to full sentences or texts.

Due to the different scopes in experimental research and NLP, it is seldom possible to directly draw conclusions concerning features from these studies to NLP: Speaking in broad terms, psycholinguistic and neurolinguistic studies provide evidence of human cognitive processing of text or speech primarily through controlled experiments. The experiment as well as the textual stimulus are carefully designed in order to isolate a specific cognitive process. Data-driven NLP works towards enabling computers to understand and manipulate naturally-occurring human language through machine learning models based on huge corpora. The phenomena that NLP models aim to model are typically much broader and less well-defined than what is examined in psycholinguistic studies.

Recently, it has become more common to implement naturalistic reading experiments (Hamilton and Huth, 2018). Naturalistic reading denotes self-paced reading of naturally-occurring text without any specific task or reading constraints, such as limiting the preview of the following

words. This allows subjects to read at their own speed and results in different reading times between subjects, which calls for more elaborate pre-processing. Naturalistic reading studies diverge from tightly controlled experimental designs and allow the participants to read continuous stimuli, i.e. full sentences or paragraphs spanning multiple lines on the screen. In addition to the more natural setting, a big advantage is the possibility to study linguistic phenomena on different levels (e.g., phonemes, syllables, words, phrases, sentences, discourse), which unfold at different timescales in the same naturalistic stimulus such as a story. Moreover, naturalistic experimental designs, which use language within the rich context of stories, audiobooks, and dialogues, produce results which are more easily generalizable to everyday language use (Kandylaki and Bornkessel-Schlesewsky, 2019). Since generalizability of results is one of the main objectives in experimental science, the potential importance of increased ecological validity in naturalistic experiment paradigms is undeniable.

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

Multilingual neurolinguistics

The majority of research in NLP, as well as most of the available cognitive data sources is in English. However, it is well known that language processing between native and foreign language speakers differs in the active brain regions (Perani et al., 1996). Moreover, second language learners exhibit different reading patterns than native speakers (Dussias, 2010).

Eye-tracking and fMRI studies on bilingualism suggest that, although the same general structures are active for both languages, differences within these general structures are present across languages and across levels of processing (Marian et al., 2003; Dehghani et al., 2017). In an effort to promote eye-tracking research of bilingual reading, Cop et al. (2017) provide an English-Dutch eye-tracking corpus tailored to analyze the bilingual reading process.

Further, there are even differences in the processing of dialects and standard variations, e.g., Lundquist and Vangsnes (2018) for Norwegian dialects and Stocker and Hartmann (2019) for variations of German. Hence, it is not only important to take language-specific aspects into account in the NLP methods, but it is crucial to account for these differences in human language processing. It remains an open questions how many of the referenced studies in this paper would generalize to other languages.

3.3. Extracting features

This section covers different approaches to find the most meaningful features from human language processing recordings.

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

Most studies combine linguistic features with gaze features (e.g., Rohanian et al. (2017) and Yaneva et al. (2018)). Further, Barrett et al. (2016a) use word frequency and word length features in combination with eye-tracking features, because the two properties explain much of the variance in fixation duration (Just and Carpenter, 1980; Levy, 2008). Results by Demberg and Keller (2008) and Lopopolo et al. (2019) showed a relation between regression features and the syntactic structure of sentences: About 40% of regressions land on target words engaged in dependency relations. Moreover, many other properties such as transitional probabilities or age of acquisition could also be used. In Hollenstein and Zhang (2019) and Barrett et al. (2018b), gaze features are combined with pre-trained word embeddings to improve performance.

All these works, however, rely on rather heavy feature engineering. Contrariwise, these features can also be predicted from text: Hahn and Keller (2016) presented an unsupervised neural model of human reading by predicting the fixations within sentences. Similarly, Matthies and Søgaard (2013) predict skipping probabilities across multiple readers. Moreover, Singh et al. (2016) introduced a method where eye movements are learned in order to alleviate the need to get the task data annotated with eye movements. A similar approach is also used by Long et al. (2019). Comparably, fMRI signals have been predicted from language model representations, e.g., Rodrigues et al. (2018) and Abnar et al. (2018).

In general, feature engineering for M/EEG and fMRI data is more a matter of dimensionality reduction. For instance, most studies leveraging M/EEG data for NLP average the signals over all electrodes or sensors (e.g., Wehbe et al. (2014b)). Moreover, methods such as principal component analysis are often used to reduce the dimensions of both M/EEG and fMRI data. In the case of fMRI data, we mention several strategies for voxel selection in Section 2.3. to reduce the number of dimensions. For M/EEG signals, it is also possible to work with frequency band features or ERPs based on neurolinguistic findings (see Section 2.2.). However, these features have not yet been explored in detail to improve NLP tasks.

Aggregating features

Controlled psycholinguistic studies include multiple subjects to obtain significant differences considering the effect

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

sizes of interest (Vasishth et al., 2018). In many NLP studies that use eye movements as word representations, eye movement metrics are averaged over several readers arguing for more stability and less noise, but most studies are limited by number of words and readers in the provided corpora (Rohanian et al., 2017; Yaneva et al., 2018; Mishra et al., 2017b; Hollenstein et al., 2019a). But how many subjects are required to obtain a robust average signal for NLP? Gaze annotation can never be a gold annotation, irrespective of the number of readers. It is intrinsically noisy and there is no uniquely correct reading pattern. Skilled readers will exhibit a more idiosyncratic reading behaviour under similar conditions. Language learners or readers with reading impairments will exhibit a noisier signal, that is difficult to use in NLP (Bingel et al., 2018). Takmaz et al. (2019) compared aggregated gaze features and sequential features for generating image captions.

Hollenstein et al. (2019a) used eye movement and EEG features to improve named entity recognition, relation classification and sentiment classification. They showed that averaging over ten skilled native readers is able to diminish the noise and variability between subjects, to the extent where the average worked almost as good as the best individual reader, for both gaze and EEG models. While subject variability is even larger in fMRI signals, averaging over participants can help to avoid overfitting (Bingel et al., 2016). Moreover, Schwartz et al. (2019) showed how a language model fine-tuned with fMRI brain activity data transfers across multiple participants.

Word-level signals

In some studies, averages of gaze features over word types have been used to alleviate the need of having gaze data at test time, and even achieved better results than token-level features (Barrett et al., 2016a; Hollenstein and Zhang, 2019). Klerke and Plank (2019) analyzed this in detail for PoS tagging and found that content words are especially sensitive to type-level gaze features.

For recordings of continuous stimuli, the EEG samples have to be mapped to the points in time where a word (or phrase) was heard or read. Hauk and Pulvermüller (2004) presented evidence that lexical access from written word stimuli is an early process that follows stimulus presentation by less than 200 ms. Between 200-500ms, the word's semantic properties are processed (Wehbe et al., 2014b). Moreover, Dimigen et al. (2011) studied the linguistic effects of eye movements and EEG signal co-registration in natural reading and showed that they accurately represent lexical processing. This suggest that, in the case of reading, the brain processes words when they are fixated for the first time, so that by mapping the EEG samples to the corresponding reading times it is possible to extract word-level EEG features. In combination with the eye-tracking, the high sampling rate of EEG allows us to get a definable signal for each token. In case of listening, the EEG signals can simply be mapped to the timestamps of the utterances. Analogous to the type aggregation approach described for eye-tracking signals, token-level EEG and fMRI features can be aggregated on word type level (Hollenstein et al., 2019a; Bingel et al., 2016). This eliminates the need of recorded data at test time, however the results are more promising for eye-tracking data than for brain activity.

In the case of fMRI, however, extracting token-level or type-level signals from continuous stimuli is less recommendable. A few studies have extracted token-level features from scans of a few seconds of duration. Bingel et al. (2016) computed individual word features for PoS induction by accounting for the hemodynamic delay using a Gaussian sliding window over a certain time window. Hollenstein et al. (2019b) also account for this delay when extraction word-level features, and then average the word features over multiple trials from different contexts. It is difficult to quantify how much of the information of single word processing is captured in these signals. In fMRI studies, models are most often trained separately for each subject due to the large individual differences. It is, however, also possible to learn a shared representation between subjects (Vodrahalli et al., 2018). Additionally, the signal can be averaged if multiple trials are available per stimulus as in Mitchell et al. (2008).

3.4. Including the features in the models

This section describes the most common machine learning methods for leveraging human cognitive processing for NLP. In most applications of systems using human data, it is sub-optimal to require real-time human features at test time. For eye-tracking, there are several studies working towards not requiring recordings during inference. We start by outlining those methods and move to other cognitive signals thereafter.

When using human language processing data recorded from continuous stimuli, it is intuitive to implement sequence labelling or sequence classification approaches. For instance, Strzyz et al. (2019) argue in favor of using bidirectional LSTMs for predicting eye-movement information. Many of other studies have leveraged similar neural architectures, for example, Klerke et al. (2016) and Hollenstein and Zhang (2019).

A basic approach is to include cognitive features as multi-dimensional vectors to represent each word, possibly along with other word-based features. For instace, Rohanian et al. (2017), Barrett and Søgaard (2015) and Yaneva et al. (2018) implemented this approach for eye-tracking data. However, this requires gaze data at test time. Barrett et al. (2016a) and Barrett et al. (2016b) showed that word-type averages of gaze features yielded better results for PoS induction than token-level features. In this case, gaze representations are used similarly to word embeddings, with which they can also be combined (Barrett et al., 2018b). Klerke and Plank (2019) analyzed this in detail for PoS tagging and showed that word type variance was better than individual gaze representations and less aggregated gaze features. Additionally, Hollenstein and Zhang (2019) showed the same advantages of type-level aggregated features for improving named entity recognition on corpora with no available gaze features during training *and* testing. However, type aggregation on EEG data has not shown the same positive benefits (Hollenstein et al., 2019a).

Concatenating cognitive features has also been tested with brain activity data. Bingel et al. (2016) concatenate ex-

tracted fMRI vectors from multiple subjects with linguistic features. Moreover, Schwartz et al. (2019) include fMRI and MEG data to augment a language model by fine-tuning a model trained on textual input with brain activity signals. In addition, multi-task learning is a method of training a system that inherently does not need human data on the test set. Multi-task learning studies typically use only one feature, but that is most likely due to constraints in the model architecture, i.e. an increasing number of parameters leading to longer training times. Hollenstein et al. (2019a) trained multi-task learning models to learn eye-tracking and EEG features at the same time as NLP tasks such as sentiment analysis and relation detection. Multi-task learning has also been successful when generalizing across subjects from EEG data, for applications such as brain-computer interfaces (Alamgir et al., 2010). Leveraging eye-tracking data, González-Garduño and Søgaard (2017), Klerke et al. (2016) and Klerke and Plank (2019) employ a multi-task learning setup for text compression, readability prediction, and syntactic tagging, respectively, while also learning to predict a gaze feature as an auxiliary task.

Lastly, another related option is to regularise the attention of a recurrent neural network with human data for sequence classification. Attention weights influence the relative importance of each word on the model, but require large amounts of data to be trained. Barrett et al. (2018a) used sentences from the main dataset to update the model parameters, while sentences from a smaller, non-overlapping eye-tracking corpus were used to only train the attention function. Regularising the attention function could also be done using other human measures such as EEG.

3.5. Measuring improvements

On one hand, natural language understanding models are mostly optimized for performance on specific tasks and typically do not transfer well to other tasks or even other datasets (Talman and Chatzikyriakidis, 2019). On the other hand, cognitive signals are typically constrained to their experimental design and stimuli. These discrepancies may lead to limitations in the possible improvements when leveraging cognitive signals to enhance NLP models.

Indeed, the improvements achieved with cognitive signals are often modest. Therefore, we want to highlight the importance of robust baselines and proper significance testing. Examples of strong baselines are, for instance, word frequency for eye-tracking signals to ensure that the cognitive features add more to the model than purely lexical aspects; or comparing EEG and fMRI feature vectors to random vectors to guarantee that the cognitive features contain more than added dimensions of noise. Additionally, after achieving better results than strong baseline models, one needs to ascertain that the improvements are not due to some artifacts in the cognitive data. Hence, it is vital to perform suitable significance tests, such as permutation tests (Dror et al., 2018).

Furthermore, Gauthier and Ivanova (2018) propose three highly sensible strategies for making language decoding studies from brain activity more interpretable: (1) committing to a specific mechanism and task, which would help to distinctly link brain activity features to specific NLP tasks, (2) dividing the input feature space into subsets that capture representations optimized for a particular task, and (3) explicitly measuring explained variance to evaluate the extent to which each model component explain the overall brain responses.

4. Ethical considerations

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

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

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

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

Acknowledgements

Author L. Beinborn was funded by the Netherlands Organisation for Scientific Research, through a Gravitation Grant 024.001.006 to the Language in Interaction Consortium.

Bibliographical References

Abnar, S., Ahmed, R., Mijnheer, M., and Zuidema, W. (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.

Abnar, S., Beinborn, L., Choenni, R., and Zuidema, W. (2019). Blackbox meets blackbox: Representational similarity & stability analysis of neural language models and brains. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 191–203.

Aggarwal, C. C., Hinneburg, A., and Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In Jan Van den Bussche et al., editors, *Database Theory — ICDT 2001*, pages 420–434, Berlin, Heidelberg. Springer Berlin Heidelberg.

Alamgir, M., Grosse-Wentrup, M., and Altun, Y. (2010). Multitask learning for brain-computer interfaces. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 17–24.

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

Anderson, A. J., Kiela, D., Clark, S., and Poesio, M. (2017). Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns. *Transactions of the Association for Computational Linguistics*, 5:17–30.

Barrett, M. and Søgaard, A. (2015). Reading behavior predicts syntactic categories. In *Proceedings of the nineteenth conference on computational natural language learning*, pages 345–249.

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

Barrett, M., Keller, F., and Søgaard, A. (2016b). Cross-lingual transfer of correlations between parts of speech and gaze features. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1330–1339.

Barrett, M., Bingel, J., Hollenstein, N., Rei, M., and Søgaard, A. (2018a). Sequence classification with human attention. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 302–312.

Barrett, M., González-Garduño, A. V., Frermann, L., and Søgaard, A. (2018b). Unsupervised induction of linguistic categories with records of reading, speaking, and writing. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2028–2038.

Beinborn, L., Abnar, S., and Choenni, R. (2019). Robust evaluation of language-brain encoding experiments. *arXiv preprint arXiv:1904.02547*.

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

Bingel, J., Barrett, M., and Søgaard, A. (2016). Extracting token-level signals of syntactic processing from fMRI - with an application to PoS induction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 747–755. ACL.

Bingel, J., Barrett, M., and Klerke, S. (2018). Predicting misreadings from gaze in children with reading difficulties. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 24–34.

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

Brennan, J. (2016). Naturalistic sentence comprehension in the brain. *Language and Linguistics Compass*, 10(7):299–313.

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

Brouwer, H., Crocker, M. W., Venhuizen, N. J., and Hoeks, J. C. (2017). A neurocomputational model of the n400 and the p600 in language processing. *Cognitive science*, 41:1318–1352.

Bulat, L., Clark, S., and Shutova, E. (2017). Speaking, seeing, understanding: Correlating semantic models with conceptual representation in the brain. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1081–1091. Association for Computational Linguistics.

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

Chen, Y.-N., Chang, K.-M., and Mostow, J. (2012). Towards using EEG to improve ASR accuracy. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 382–385. Association for Computational Linguistics.

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

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

Craik, A., He, Y., and Contreras-Vidal, J. L. (2019). Deep learning for electroencephalogram (EEG) classifi-

cation tasks: a review. *Journal of neural engineering*, 16(3):031001.

Dalmaijer, E. (2014). Is the low-cost EyeTribe eye tracker any good for research? Technical report, PeerJ PrePrints.

Dehghani, M., Boghrati, R., Man, K., Hoover, J., Gimbel, S. I., Vaswani, A., Zevin, J. D., Immordino-Yang, M. H., Gordon, A. S., Damasio, A., et al. (2017). Decoding the neural representation of story meanings across languages. *Human brain mapping*, 38(12):6096–6106.

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

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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.

Dimigen, O., Sommer, W., Hohlfeld, A., Jacobs, A. M., and Kliegl, R. (2011). Coregistration of eye movements and EEG in natural reading: analyses and review. *Journal of Experimental Psychology: General*, 140(4):552.

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

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

Ehinger, B. V. and Dimigen, O. (2019). Unfold: an integrated toolbox for overlap correction, non-linear modeling, and regression-based eeg analysis. *PeerJ*, 7:e7838.

Eugster, M. J., Ruotsalo, T., Spapé, M. M., Kosunen, I., Barral, O., Ravaja, N., Jacucci, G., and Kaski, S. (2014). Predicting term-relevance from brain signals. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 425–434. ACM.

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

Faruqui, M. and Dyer, C. (2014). Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471.

Fedorenko, E., Hsieh, P.-J., Nieto-Castañón, A., Whitfield-Gabrieli, S., and Kanwisher, N. (2010). New method for fMRI investigations of language: defining ROIs functionally in individual subjects. *Journal of neurophysiology*, 104(2):1177–1194.

Frank, S. L., Monsalve, I. F., Thompson, R. L., and Vigliocco, G. (2013). Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods*, 45(4):1182–1190.

Frank, S. L., Otten, L. J., Galli, G., and Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and language*, 140:1–11.

Funke, G., Greenlee, E., Carter, M., Dukes, A., Brown, R., and Menke, L. (2016). Which eye tracker is right for your research? Performance evaluation of several cost variant eye trackers. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 60, pages 1240–1244. SAGE Publications Sage CA: Los Angeles, CA.

Futrell, R., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S., and Fedorenko, E. (2018). The Natural Stories Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Fyshe, A., Talukdar, P. P., Murphy, B., and Mitchell, T. M. (2014). Interpretable semantic vectors from a joint model of brain-and text-based meaning. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2014, page 489. NIH Public Access.

Gauthier, J. and Ivanova, A. (2018). Does the brain represent words? An evaluation of brain decoding studies of language understanding. *arXiv preprint arXiv:1806.00591*.

Gauthier, J. and Levy, R. (2019). Linking artificial and human neural representations of language. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 529–539.

Gibaldi, A., Vanegas, M., Bex, P. J., and Maiello, G. (2017). Evaluation of the Tobii EyeX eye tracking controller and Matlab toolkit for research. *Behavior research methods*, 49(3):923–946.

González-Garduño, A. V. and Søgaard, A. (2017). Using gaze to predict text readability. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 438–443.

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

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

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

Henderson, J. M., Luke, S. G., Schmidt, J., and Richards, J. E. (2013). Co-registration of eye movements and event-related potentials in connected-text paragraph reading. *Frontiers in systems neuroscience*, 7:28.

Henderson, J. M., Choi, W., Luke, S. G., and Desai, R. H. (2015). Neural correlates of fixation duration in natural reading: Evidence from fixation-related fMRI. *Neu-*

roImage, 119:390–397.

Henderson, J. M., Choi, W., Lowder, M. W., and Ferreira, F. (2016). Language structure in the brain: A fixation-related fMRI study of syntactic surprisal in reading. *Neuroimage*, 132:293–300.

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83.

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

Hollenstein, N. and Zhang, C. (2019). Entity recognition at first sight: Improving NER with eye movement information. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

Hollenstein, N., Rotsztejn, J., Troendle, M., Pedroni, A., Zhang, C., and Langer, N. (2018). ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Scientific Data*.

Hollenstein, N., Barrett, M., Troendle, M., Bigiolli, F., Langer, N., and Zhang, C. (2019a). Advancing NLP with cognitive language processing signals. In *arXiv preprint arXiv:1904.02682*.

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

Hollenstein, N., Troendle, M., Zhang, C., and Langer, N. (2019c). Zuco 2.0: A dataset of physiological recordings during natural reading and annotation. *arXiv preprint arXiv:1912.00903*.

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

Huang, H., Hu, X., Zhao, Y., Makkie, M., Dong, Q., Zhao, S., Guo, L., and Liu, T. (2017). Modeling task fmri data via deep convolutional autoencoder. *IEEE transactions on medical imaging*, 37(7):1551–1561.

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

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

Kandylaki, K. D. and Bornkessel-Schlesewsky, I. (2019). From story comprehension to the neurobiology of language.

Klerke, S. and Plank, B. (2019). At a glance: The impact of gaze aggregation views on syntactic tagging. In *Proceedings of the Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)*, pages 51–61.

Klerke, S., Goldberg, Y., and Søgaard, A. (2016). Improving sentence compression by learning to predict gaze. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1528–1533.

Klerke, S., Madsen, J. A., Jacobsen, E. J., and Hansen, J. P. (2018). Substantiating reading teachers with scanpaths. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, pages 1–3.

Klimesch, W. (2012). Alpha-band oscillations, attention, and controlled access to stored information. *Trends in cognitive sciences*, 16(12):606–617.

Kriegeskorte, N., Goebel, R., and Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, 103(10):3863–3868.

Kunze, K., Kawaichi, H., Yoshimura, K., and Kise, K. (2013). Towards inferring language expertise using eye tracking. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pages 217–222. ACM.

Kutas, M. and Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in cognitive sciences*, 4(12):463–470.

LaConte, S., Anderson, J., Muley, S., Ashe, J., Frutiger, S., Rehm, K., Hansen, L. K., Yacoub, E., Hu, X., Rottenberg, D., et al. (2003). The evaluation of preprocessing choices in single-subject bold fmri using npairs performance metrics. *NeuroImage*, 18(1):10–27.

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

Li, M. and Lu, B.-L. (2009). Emotion classification based on gamma-band EEG. In *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, pages 1223–1226. IEEE.

Li, J., Fabre, M., Luh, W.-M., and Hale, J. (2018). The role of syntax during pronoun resolution: Evidence from fMRI. In *Proceedings of the Eight Workshop on Cognitive Aspects of Computational Language Learning and Processing*, pages 56–64.

Long, Y., Xiang, R., Lu, Q., Huang, C.-R., and Li, M. (2019). Improving attention model based on cognition grounded data for sentiment analysis. *IEEE Transactions on Affective Computing*.

Lopopolo, A., Frank, S. L., van den Bosch, A., and Willems, R. (2019). Dependency parsing with your eyes: Dependency structure predicts eye regressions during reading. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 77–85.

Lundquist, B. and Vangsnes, Ø. A. (2018). Language separation in bidialectal speakers: Evidence from eye tracking. *Frontiers in psychology*, 9:1394.

Marian, V., Spivey, M., and Hirsch, J. (2003). Shared and separate systems in bilingual language processing: Converging evidence from eyetracking and brain imaging. *Brain and language*, 86(1):70–82.

Matthies, F. and Søgaard, A. (2013). With blinkers on: Robust prediction of eye movements across readers. *Proceedings of the 2013 Conference on empirical methods in natural language processing (EMNLP)*, pages 803–807.

Mensch, A., Mairal, J., Bzdok, D., Thirion, B., and Varoquaux, G. (2017). Learning neural representations of

human cognition across many fMRI studies. In *Advances in Neural Information Processing Systems*, pages 5883–5893.

Michel, V., Gramfort, A., Varoquaux, G., Eger, E., and Thirion, B. (2011). Total variation regularization for fmri-based prediction of behavior. *IEEE transactions on medical imaging*, 30(7):1328–1340.

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

Minnema, G. and Herbelot, A. (2019). From brain space to distributional space: the perilous journeys of fMRI decoding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 155–161.

Mishra, A., Kanojia, D., Nagar, S., Dey, K., and Bhattacharyya, P. (2016). Harnessing cognitive features for sarcasm detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1095–1104.

Mishra, A., Dey, K., and Bhattacharyya, P. (2017a). Learning cognitive features from gaze data for sentiment and sarcasm classification using convolutional neural network. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 377–387. ACL.

Mishra, A., Kanojia, D., Nagar, S., Dey, K., and Bhattacharyya, P. (2017b). Leveraging cognitive features for sentiment analysis. *Proceedings of The 20th Conference on Computational Natural Language Learning*, pages 156–166.

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

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

Murphy, B. and Poesio, M. (2010). Detecting semantic category in simultaneous EEG/MEG recordings. In *Proceedings of the NAACL HLT 2010 first workshop on computational neurolinguistics*, pages 36–44. Association for Computational Linguistics.

Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., and Hays, J. (2016). WebGazer: Scalable webcam eye tracking using user interactions. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence-IJCAI 2016*.

Parthasarathy, S. and Busso, C. (2017). Jointly predicting arousal, valence and dominance with multi-task learning. In *Interspeech*, pages 1103–1107.

Pedroni, A., Bahreini, A., and Langer, N. (2019). Automagic: Standardized preprocessing of big EEG data. *NeuroImage*.

Perani, D., Dehaene, S., Grassi, F., Cohen, L., Cappa, S. F., Dupoux, E., Fazio, F., and Mehler, J. (1996). Brain processing of native and foreign languages. *NeuroReport-International Journal for Rapid Communications of Research in Neuroscience*, 7(15):2439–2444.

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., Botvinick, M., and Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nature communications*, 9(1):963.

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

Rodrigues, J. A., Branco, R., Silva, J., Saedi, C., and Branco, A. (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.

Rohanian, O., Taslimipoor, S., Yaneva, V., and Ha, L. A. (2017). Using gaze data to predict multiword expressions. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 601–609.

Rowtula, V., Oota, S., Gupta, M., and Surampudi, B. R. (2018). A deep autoencoder for near-perfect fMRI encoding. In *Workshop on Modeling and Decision-Making in the Spatiotemporal Domain, 32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*.

Schwartz, D., Toneva, M., and Wehbe, L. (2019). Inducing brain-relevant bias in natural language processing models. In *Advances in Neural Information Processing Systems*, pages 14100–14110.

Shain, C., Blank, I. A., van Schijndel, M., Schuler, W., and Fedorenko, E. (2019). fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, page 107307.

Singh, A. D., Mehta, P., Husain, S., and Rajakrishnan, R. (2016). Quantifying sentence complexity based on eye-tracking measures. In *Proceedings of the workshop on computational linguistics for linguistic complexity (cl4lc)*, pages 202–212.

Søgaard, A. (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.

Stocker, K. and Hartmann, M. (2019). "Next Wednesday's meeting has been moved forward two days": The time-perspective question is ambiguous in Swiss German, but not in Standard German. *Swiss Journal of Psychology*, 78(1-2):61.

Strzyz, M., Vilares, D., and Gómez-Rodríguez, C. (2019). Towards making a dependency parser see. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1500–1506.

Suster, S., Tulkens, S., and Daelemans, W. (2017). A short review of ethical challenges in clinical natural language processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 80–87.

Takmaz, E., Beinborn, L., Pezzelle, S., and Fernández, R.

(2019). Enhancing image captioning with eye-tracking. In *EurNLP*.

Talman, A. and Chatzikyriakidis, S. (2019). Testing the generalization power of neural network models across NLI benchmarks. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 85–94.

Toneva, M. and Wehbe, L. (2019). Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Advances in Neural Information Processing Systems*, pages 14928–14938.

Vasishth, S., Mertzen, D., Jäger, L. A., and Gelman, A. (2018). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language*, 103:151–175.

Vodrahalli, K., Chen, P.-H., Liang, Y., Baldassano, C., Chen, J., Yong, E., Honey, C., Hasson, U., Ramadge, P., Norman, K. A., et al. (2018). Mapping between fMRI responses to movies and their natural language annotations. *Neuroimage*, 180:223–231.

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

Wallot, S., O'Brien, B., Coey, C. A., and Kelty-Stephen, D. (2015). Power-law fluctuations in eye movements predict text comprehension during connected text reading. In *CogSci*.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2019). Glue: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019*.

Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., and Mitchell, T. (2014a). Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one*, 9(11):e112575.

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

Wikibooks. (2020). Neuroimaging data processing — Wikibooks, the free textbook project. [Online; accessed 21-February-2020].

Williams, C. C., Kappen, M., Hassall, C. D., Wright, B., and Krigolson, O. E. (2019). Thinking theta and alpha: Mechanisms of intuitive and analytical reasoning. *NeuroImage*, 189:574–580.

Wu, E. X. W., Laeng, B., and Magnussen, S. (2012). Through the eyes of the own-race bias: Eye-tracking and pupillometry during face recognition. *Social neuroscience*, 7(2):202–216.

Xu, S., Jiang, H., and Lau, F. (2009). User-oriented document summarization through vision-based eye-tracking. In *Proceedings of the 14th international conference on Intelligent user interfaces*, pages 7–16. ACM.

Yaneva, V., Evans, R., Mitkov, R., et al. (2018). Classifying referential and non-referential it using gaze. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4896–4901.

Zhang, X., Yao, L., Sheng, Q. Z., Kanhere, S. S., Gu, T., and Zhang, D. (2018). Converting your thoughts to texts: Enabling brain typing via deep feature learning of eeg signals. In *2018 IEEE international conference on pervasive computing and communications (PerCom)*, pages 1–10. IEEE.