Using EEG to Decode Semantics During an Artificial Language Learning Task

by

Chris Foster
BCS, Thompson Rivers University, 2016

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Computer Science

Using EEG to Decode Semantics
During an Artificial Language Learning Task

by

Chris Foster
BCS, Thompson Rivers University, 2016

Supervisory Committee

Dr. Alona Fyshe, Supervisor
Department of Computer Science

Dr. George Tzanetakis, Departmental Member
Department of Computer Science

ABSTRACT

The study of semantics in the brain explores how the brain represents, processes, and learns the meaning of language. In this thesis we show both that semantic representations can be decoded from electroencephalography data, and that we can detect the emergence of semantic representations as participants learn an artificial language mapping. We collected electroencephalography data while participants performed a reinforcement learning task that simulates learning an artificial language, and then developed a machine learning semantic representation model to predict semantics as a word-to-symbol mapping was learned. Our results show that 1) we can detect a reward positivity when participants correctly identify a symbol's meaning; 2) the reward positivity diminishes for subsequent correct trials; 3) we can detect neural correlates of the semantic mapping as it is formed; and 4) the localization of the neural representations is heavily distributed. Our work shows that language learning can be monitored using EEG, and that the semantics of even newly-learned word mappings can be detected using EEG.

Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

I would like to thank:

Dr. Alona Fyshe, for providing me with mentorship, support, and knowledge.

my friends, colleagues, and family, for everything family and friends can do.

Chad C. Williams, for sharing his EEG expertise and assisting with this research.

Compute Canada, for providing the compute resources used in this research.

NSERC Canada, for funding this research with the CGSM program.

DEDICATION

To Tatianna, my partner and friend.
Thank you for all you do for me.

Chapter 1

Introduction

Each year millions of people will study a foreign language. At first, the foreign symbols have no meaning. It is only through dedication and practice that these symbols become interpretable concepts. What is happening in the brain when we learn a mapping from one language to another? Can we better understand this process by measuring waves from the brain? Although the neural representations of words have been studied extensively with native languages, they remain comparatively unexplored during the learning of new languages.

Semantics is the branch of linguistics concerned with the study of *meaning*. Semanticists study the connection between the symbols, words, signs, and phrases we use in language and the conceptual idea that those signifiers represent [16]. The study of semantics in the brain explores how the brain represents, processes, and learns these semantics. These functions are argued to be critical to human cognition and communication across all languages and cultures [7].

In this thesis we show that semantic representations of the native word (e.g., "book") can be decoded from electrophysiological data measured in the human brain using machine learning and a technology known as Electroencephalography (EEG). In the past, this has been done using two other technologies for measuring brain activity: functional Magnetic Resonance Imaging (fMRI) and Magnetoencephalography (MEG). Our approach, using EEG, has a number of benefits. Particularly, it is often several levels of magnitude cheaper than other technologies and generally requires less training to operate.

We also show that, using a similar approach, we can detect the process of learning by monitoring how the semantic representations of the native word develop during an artificial language learning task. To our knowledge, this is the first application of

this methodology to research learning in the brain. Traditionally, this is done with a technique which measures a specific brain response known as the *reward positivity*. Our approach has a number of benefits over traditional measurements of learning, including the ability to measure the retention of knowledge and the speed at which learning occurs.

1.1 Thesis Overview

The rest of this chapter will introduce the important background concepts needed, and detail our specific contributions to the state-of-the-art. The chapter following reviews the key literature which we build on in this thesis. Chapter 3 introduces the methodologies used in this thesis, including the data collection paradigm, participant information, preprocessing of EEG data, and the computational model for analyzing the semantic representations. With the methodologies in mind, Chapter 4 covers the six experiments that we performed using our model and Chapter 5 includes the results of those experiments. In Chapter 6, we discuss the conclusions that we can draw from our results and how the results align with existing literature. The last chapter summarizes our new contributions to the field and concludes the thesis.

This research expands on previous work by adapting the existing semantic analysis methodology [27, 38] to EEG data. We use a novel experimental design, in which participants perform a reinforcement learning task that guides them through learning an artificial language. To detect semantics we trained a machine learning model to map from raw EEG signals to word vectors derived from an artificial neural network. We evaluate this model using the 2 vs. 2 test, a method originally developed by Mitchell et al. [27], that simplifies a complex multivariate regression model into a binary classification task. The 2 vs. 2 test is done by performing a "leave two out" cross validation in which the two left out vectors are predicted and compared to determine if the predicted vectors are closer to their own ground truth than they are to the other vector's ground truth. The percentage of predictions which are closer to their own ground truth provides a measurement of the correlation between the brain data and the word vectors.

In this thesis we show that we can: 1) detect the semantic representation of English words in EEG while participants read the symbol language, 2) measure how these semantic representations develop over time during the participant learning phase, 3) validate and compare our model against a traditional reward positivity analysis of

the same experiment, 4) provide supportive evidence that suggests intuitive alignment with the participants' task accuracies, 5) identify a delayed peak in the strength of the semantic representation correlating to the delay required for the task, and 6) provide further evidence that the source of semantic representations in the human brain is highly distributed and not simply attributable to a single area of the cortex.

1.2 Background and Terminology

This section will briefly introduce the relevant terminology at a high level, for those not familiar with the fields of neuroscience or machine learning. With a cursory understanding of these, the topics discussed in the thesis should be easier to follow. This is not designed to be a complete review of the topics and readers are encouraged to investigate further if a topic is unfamiliar to them.

Machine learning is a subfield of artificial intelligence that gives computers the ability to improve their performance at a task in response to data about that task (called training data). This is done using statistical techniques. The most common type of machine learning is supervised learning, in which training data consists of pairs of inputs and desired outputs. The algorithm learns to map from a given input x to a predicted output $\hat{y}$ using training data sets with inputs X and matching outputs Y.

In this work we use linear ridge regression, a type of machine learning algorithm. Regression indicates that the model predicts a scalar value, rather than a category. We use sets of these to make a multi-output prediction, known as a multivariate linear regression. A linear model is a model which learns a polynomial function with a degree of at most one (that is, it predicts in a straight line). Ridge regression, also known as weight decay, is a type of linear regression that utilizes a regularization mechanism. This is designed to help the algorithm perform better on results which are not in the training set (known as generalization). Regularization is especially important when there are many irrelevant features in the data matrix X. A linear model takes a set of inputs $x_1, x_2, ...x_n$ and predicts an output $\hat{y}$ by multiplying weights $w_1, w_2, ...w_n$ with the inputs and adding the components together. The weights for a linear model can be found both with closed-form expressions such as in a Cholesky least squares solver or with iterative methods such as stochastic gradient descent, though a detailed survey of methods for weight estimation are outside of the scope of this summary.

Another topic that is discussed are artificial neural networks. These are computa-

tional systems which are vaguely inspired by the connections in the brain. A *neuron* is a single unit which receives inputs x and applies weights w to each input respectively (similar to linear models). However, after summing the components a special function known as an activation function is applied to the result. The result of the activation function is the output of the neural. The activation function is typically a non-linear function, which allows the artificial neural network to learn to model non-linear data. As the phrase "network" suggests, these neurons are generally connected in series and parallel to form multiple layers of computation. The weights for a neural network are found using a process known as *gradient descent*.

Some neuroscience terminology is utilized here as well. The *cortex* refers to the outermost layer of an organ in the body. In all cases here, we are referring to the cerebral cortex, which is the outermost layer of the cerebrum. The cerebrum is the upper, largest section of the human brain which is associated with higher brain operations such as speech, movement, sensory processing, and other functions. Much of this functionality resides directly on the cortex. The cortex is categorized into four lobes: the frontal, parietal, temporal, and occipital lobes. Different methods of measuring activity in the brain identify signals better from some areas and components of the brain than from others (covered in Section 1.3), and we will also reference these areas when discussing source localization.

Electrophysiology is the study of electrical activity in biological tissues. In neuroscience, the electrophysiology signals of interest are the electrical signals from the nervous system of the body. Specifically, we are interested in measuring the electrical signals created when neurons of the brain fire in a coordinated fashion.

As mentioned prior, semantics refers to the study of meaning. This research area explores how the brain represents these semantics. When we discuss the semantic representation of a word in the brain, we are referring to the electrophysiological state of the brain while it is processing the meaning of a given word.

1.3 Collection Methodologies

In this work we commonly reference three methods of measuring electrophysiology activity in different areas of the brain: EEG, fMRI, and MEG. In this section we will discuss the measurement function for each, as well as compare the benefits and drawbacks between them. EEG is the collection methodology used in our work, but a baseline understanding of fMRI and MEG is valuable for understanding how our

research fits into the state-of-the-art and for comparing the results across collection methodologies.

All three of these methods are referring to *noninvasive* collection, which means that they do not require incision into the body to be used. Invasive sensors are used for some research, as they can provide a clearer signal or capture a smaller selection of neurons than these methods, but are generally only used for research on non-human participants. As our research is related to the understanding of semantics and language, noninvasive sensors on human participants are more common.

EEG measures the electrical activity generated by the firing of very large groups of neurons in the brain. To do this, electrodes are placed on the scalp of the participant with a conductive gel applied. The EEG voltage signal represents a difference between the voltages at two electrodes, generally the source electrode and an electrode that is identified as the *reference* electrode.

MEG measures the magnetic fields generated by the electrical current caused by the firing of very large groups of neurons in the brain. To do this, participants put their head into a helmet-shaped opening in the MEG collection device. Collection of MEG data must be performed in a magnetically shielded room.

fMRI measures the blood-oxygen-level dependent (BOLD) contrast generated by the firing of very large groups of neurons in the brain. When neurons fire, they require sugar and oxygen to be replenished from the blood stream. This causes a measurable change in the magnetism of the blood. This effect occurs much slower than detecting the direct electrical activity of neurons firing, and it also must be performed in a magnetically shielded room.

An overview comparison of the different collection modalities can be found in Table 1.1. Because EEG and MEG both measure at the scalp of the participant, they are less capable of measuring subcortical activity in the brain than fMRI. Further, the electrical signals measured by EEG do not diffuse through the skull and scalp as well as the magnetic fields detected by MEG, making EEG more susceptible to noise. However, EEG can provide an alternative view to MEG as the two modalities respond to spherical sources in the brain differently [6]. Due to the lower cost of equipment, non-dependence on a magnetically shielded room, and reduced training requirements, EEG collection is also more cost effective than fMRI or MEG. Despite the challenges with noise, EEG has a lower barrier to research and provides a different angle on the activity in the brain, which makes it a valuable tool worth exploring for semantic representation research.

Type	Magnetic Shielding	Spatial Resolution	Temporal Resolution	Cost
EEG	Not Required	Low	High	Low
MEG	Required	Medium	High	High
fMRI	Required	High	Low	High

Table 1.1: A comparison of the different brain data collection methodologies discussed in this thesis.

1.4 Contributions

Many people were involved in the making of this research. Additionally, much of this research has been published or submitted for consideration at a publisher. This thesis will therefore include components which have been published or were done by other researchers. This section will outline the contributions of everyone and any relevant publications (pending or otherwise), so no work is misrepresented.

Chad C. Williams, Dr. Olav E. Krigolson, and others from the Neuroeconomics Lab at the University of Victoria collected the data for this thesis and performed the EEG preprocessing in coordination with us. Chad performed the reward positivity analysis on the data as well. Their EEG expertise was invaluable through many components of the project. The word vectors for this thesis were provided by wordvectors.org and trained by Mikolov et al. [24].

An early version of the work in this thesis was published at the Inaugural Conference on Cognitive Computational Neuroscience. A complete version of the work in this thesis has been submitted for consideration at the journal NeuroImage.

1.5 Conclusion

In this chapter we have introduced the high level concepts and research topics which will be explored, as well as the technologies and terminologies involved. This research shows that semantic representations of the native word can be decoded from EEG data when a person views the foreign orthographic form, once the participant has successfully learned an artificial language mapping. We use existing methods for detecting semantic representations [38], and apply them in a novel language learning paradigm. We provide supporting evidence for this method using event related po-

tential (ERP), behavioral, time, and sensor based analysis techniques. In the next chapter, we will detail the key background work that we build on in this thesis.

Chapter 2

Previous Work

2.1 Early Semantic Research

In very early work, semantics in the brain have been analyzed through the traditional detection of ERPs in EEG data [21, 20]. When participants read a sentence that involves a semantically inappropriate statement (e.g., he spread the warm bread with socks), the brain elicits a measurement ERP response known as the N400. This negative component peaks approximately 400 milliseconds after stimulus onset, hence the name N400.

While visual inspection of evoked EEG data can be useful for measuring phenomena that are directly visible in the data, such as with ERPs, more complex patterns can be detected with automated analysis methods such as machine learning techniques. This can be useful for identifying attributes of the EEG data that are not tied to simple magnitude comparisons, or when analysis needs to be performed on an online setting (i.e., in real time). Additionally, grand average ERPs can be different in timing and amplitude between participants depending on age variations [8]. For example, an early application of machine learning classification on brain data was the use by Wang et al. to detect whether or not participants were viewing a picture of reading sentences based on their fMRI activity [41].

Machine learning methods can also be useful for detecting semantic information. Mitchell et al. were able to categorize trials of participants reading a word into one of twelve semantic categories based on the word in an early paper [26]. Similarly to the research based on ERPs, they could also detect when a participant found a sentence to be semantically ambiguous. Shinkareva et al. was able to identify individual

concepts and their corresponding semantic categories for a participant based only on the training data from other participants [37]. This indicated the existence of stable semantic representations of concepts in the brain that are shared across people. While much of this previous work was done in fMRI, there was similar work using EEG that provided evidence of the ability to identify limited semantics. For example, Gu et al. was able to perform sentiment analysis (a more simple type of semantic analysis that categorizes concepts into positive, negative, or neutral categories) of a limited set of sentences using EEG data [11]. However, work in this area utilizing EEG did not quite match the level of semantic detail that was found in studies using fMRI.

2.2 Generalizable Semantic Models

Until 2008, most research utilized models that required many repeated training examples of a stimulus before it could correctly identify that stimulus in the future [21, 20, 41, 26, 37, 11]. In effect, this means that these models are only capable of recognizing brain states that the machine learning algorithm had already been trained to recognize. As neuroscience training data is very expensive to collect compared to other applications of machine learning, this could be viewed as a limitation of the semantic models. It would be impractical to collect sufficient trials of every possible word in the English language for each subject.

By training a machine learning model to accurately predict the expected fMRI activity of a participant reading concrete nouns, Mitchell et al. showed that the semantic features of a word are correlated with fMRI data of a participant viewing the word [27]. Although the model is trained using observed fMRI data of participants reading 60 concrete nouns, the model is capable of generating predictions for thousands of words for which it has never seen fMRI data. This is achieved by encoding each word as a vector of intermediate features based on the co-occurrences of the word with 25 verbs in a large text corpus. Rather than training the model to recall a given word categorization, this forces the weights to model the semantic patterns in the brain. Mitchell et al. demonstrated a direct relationship between the statistics of word co-occurrence and the neural activation associated with each word's meaning [27].

Another key study in this area reproduced Mitchell et al. [27] using MEG [38]. Sudre et al. used MEG data and word vectors to correctly identify concrete nouns. However, in this work, the word vectors were based on human responses to semantic

questions about the word (e.g. Is it alive? It it bigger than a golf ball?) rather than automatically generated features from text corpora. The use of MEG allowed Sudre et al. to pinpoint in time when the semantics of a word could be detected and when the strength of the representation was the strongest. Subsequent work showed that, with some fine tuning, word vectors derived from a text corpus could be as accurate for predicting the word a person is reading as the behavioral vectors used in Sudre et al. [31].

While most of the work discussed here focuses on the analysis of single concrete nouns, recent work has been done that extends into more complicated language structures such as phrases or sentences [5, 10, 33]. In our work, we will focus on adapting the single word paradigm to EEG. However, with this ground work established in EEG more complicated language structures become an obvious area for future experimentation.

2.3 Research in EEG

Compared to fMRI and MEG, EEG data has remained comparatively underutilized for the fine-grained distinction of individual words. This may be due to the challenges that come with EEG data (e.g. lower spatial resolution, comparatively poor signal-to-noise ratio). One of the first studies to successfully use word vectors to differentiate words EEG was performed by Murphy et al in 2009 [29]. In addition, they were able to distinguish between two semantic classes (land mammals or work tools) [29, 30]. The accuracy was as high as 98% when averaged over multiple analyses, providing evidence that EEG could give more cost effective exploration of brain-based semantics in more naturalistic environments. This thesis adds to the body of evidence that EEG can be used to model semantics representations with significant accuracy.

2.4 Learning-related Literature

In addition to studying representation, in this work we also examine learning. Our novel experimental design also allows us to study participant learning in a unique fashion by applying a machine learning model of semantic representations. Learning has been traditionally studied in EEG using ERPs. The ERP component of particular interest for learning is known as the *reward positivity* [35]. The reward positivity has also been known as the feedback error related negativity (fERN), medial frontal

negativity (MFN), feedback related negativity (RFN), or feedback negativity (FN). This signal is a robust, time-locked ERP component occurring approximately 250 ms following error feedback. It is suggested that the reward positivity reflects the activity of a generic error monitoring system in the brain [25]. It is known to be associated with win/loss processing.

The amplitude of the reward positivity is associated with behavior-measured learning when presented in a reinforcement learning paradigm such as the one we use in this thesis [14, 39, 42]. However, the exact nature of the reward positivity's association with learning remains unclear and debated. In some work the reward positivity is found to have a progressively reduced amplitude as participants perform better on the task, and in other work this correlation has not been consistently detected [40]. We aim to provide an alternative tool for analyzing learning in this paradigm, which may provide insight into the reward positivity and offer other benefits.

2.5 Conclusion

Traditional analysis methods for semantic information in the brain consist mostly of ERP-based techniques, however machine learning methods have been able to provide additional insight over magnitude based visual comparisons. Mitchell et al. built on these early methods to create an approach that utilizes semantic word vectors generated from a text corpus [27]. This approach models the actual semantics of words rather than learning a mapping between the brain data and a category, and introduces the ability to generalize to words the model has never been trained on before.

This work, original in fMRI, was further iterated on when adapted to MEG [38]. It has also been expanded to include more complex language structures such as sentences and adjective-noun phrases [5, 33, 10]. Our work builds on these to adapt the corpus-based approach from Mitchell et al. and the iterations from Sudre et al. to the EEG collection methodology and our reinforcement learning based experiment paradigm. With this new paradigm and adapted approach we hope to provide insight into the learning process of the brain, something traditionally studied by an ERP component known as the reward positivity. The following chapter will describe the experiment paradigm, preprocessing techniques, and model framework we use in our experiments.

Chapter 3

Methodologies

3.1 Introduction

Our research shows that it is possible to use EEG to track the emergence of semantic mappings in participants during an artificial language learning task. Our methodology adapts an existing semantic analysis approach based on machine learning and applies it to an EEG-based reinforcement learning experiment paradigm. This allows us to model the development of semantic mappings. In this chapter we describe the collection methodology and the task performed by the participants, followed by a detailed definition of the machine learning semantic representation model and evaluation framework. The model that we use attempts to find a mapping between the EEG data and semantic word vectors used in computational linguistics. If the model is able to detect a relationship, it will be detected by a statistical test known as the 2 vs. 2 test. Lastly, we describe the statistical methods used to validate our results and summarize the methodology.

3.2 Data Collection

We collected data for 30 participants, via an EEG monitor equipped with 64 sensors (ActiCHamp, Revision 2, Brainproducts GmbH, Munich, Germany). Five participants were excluded: two participants due to technical issues with behavioral data collection, two participants due to technical issues with EEG collection, and one participant who did not follow task instructions. The 25 remaining participants consisted of 9 males and 16 females with an average age of 20 years and average self-evaluated

Figure 3.1: The experimental paradigm. Participants were required to learn a mapping of symbols to English words through trial and error. This simulates vocabulary learning.

English fluency of 9.7 out of 10. The majority (21 of 25) were right handed. Of the 64 sensors, two mastoid sensors were dedicated as reference electrodes and another electrode was used as the ground, leaving 61 total signal electrodes. Collection was performed in a sound-dampened room with participants facing a 19" LCD screen and interacting with the experiment using a button controller (VPixx, Vision Science Solutions, Quebec, Canada). The task was written in MATLAB (Version 8.6, Mathworks, Natick, U.S.A.) using the Psychophysics Toolbox extension [4].

Participants viewed a series of symbols from the Tamil and Manipuri alphabets, which were assigned to a random English word. The randomization was consistent across all participants, but had no relationship to the meaning of the Tamil or Manipuri words. While other artificial mappings may have been utilized instead, these symbols were not likely to be familiar to candidate participants, were readily available existing components from a real language, and ensured roughly equal difficulty of translation (for other languages the translation of a word may be more obvious to an English speaker than for other words). Utilizing symbols also has interesting implications when comparing our results with prior work, in which participants viewing visual images and English words could be criticized as a detection of visual features rather than of semantics (see Section 6.1 for more details).

There were a total of 60 words and symbols, 43 of which have a definitive part of speech category. These consist of 3 pronouns, 3 verbs, 14 adjectives, and 23 abstract or concrete nouns. The remaining 17 may take the role of multiple parts of speech, for example *north* may act as either an abstract noun, adverb, or adjective and *run* may be either an abstract noun or a verb. A complete list of symbols and words is available in Appendix A.

Participants were presented with a symbol, and asked to select the correct word

from four options. The participant received visual feedback about their response: correct ("✔") or incorrect ("X"). Figure 3.1 illustrates a single trial of the paradigm. This simulates learning a language through trial and error. We hypothesized that as participants learned the mapping of symbols–to–words, they would also assign semantic meaning to each symbol. Our task used a 1–to–1 mapping of symbols–to–words over a very small subset of English. Of course, this is not representative of learning a complete language but allowed us to detect the process of learning the symbol–to–word mapping, mirroring vocabulary learning.

To facilitate learning, symbols were selected from a set that grew as the experiment progressed. During the first block, participants were presented with six symbols (representing three pronouns, three verbs). In subsequent blocks, three new symbols (and thus three new words) were added. These three new symbols were randomly paired with three previously seen symbols so that each block cycled through six symbols. There were a total of 19 blocks, and 60 total symbols learned. Throughout the experiment, each of the participants viewed a random number of trials (ranging from 0–20, denoted as n_t) for each of the 60 symbols. After the first block, the order in which symbols were added was randomly determined, so that no two participants viewed the noun symbols in the same order.

The stimuli were displayed on a gray background. Each trial begins with a black fixation cross for 700 to 1000 ms, followed by a symbol written in black, 4.5 cm^2 in size. The symbol presented was randomly selected from the list of six for the block. After 500 ms, four black English words appeared in the arrangement of a fixation cross (top, bottom, right, left) below the symbol. One of the choices was the correct answer, and the three distractor words (incorrect answers) were randomly chosen from the remaining five words. The assignment of words to the four locations was randomly determined. Participants were instructed to respond by pressing one of the buttons on the RESPONSEPixx controller, which also has response buttons arranged in a cross. Once a participant made a selection, the selected word turned white for 500 ms, the screen changed to a fixation for 700 to 1000 ms, and a feedback stimulus appeared for one second ("✔" or "X"). If a selection was not made within two seconds, an exclamation mark would appear to signify that they took too long to respond. Within a block, ten symbols were presented sequentially one at a time and then evaluated for accuracy. Participants stayed on the current block until receiving 90% or higher accuracy over the set of ten.

To further facilitate the transfer of meaning to symbols, participants also viewed

three word sentences containing one pronoun, one verb, and one noun (e.g., *I am happy*). The sentence phases displayed three sentences before and after each word learning phase described above. In these phases, participants saw one word at a time for one second each, separated by a fixation cross for 700 to 1000 ms, which was followed by four multiple choice answers as to indicate what the sentence had said. For the purposes of this thesis, the sentence trials were discarded. The participants each saw on average 667 ($\sigma = 79$) word exposures, including sentences, with breaks provided. The average task accuracy of individual participants ranges from 72% - 90% and the mean over all participants is 81%. The standard deviation of average task accuracy is 4%.

3.2.1 Data Preprocessing

EEG data generally contains artifacts that must be identified and corrected for. For example, artifacts are generated by the electrical activity of the muscular movements associated with eye blinks or even eye movements. Other movements, such as a turning the head, will also generate artifacts. Further, movement over time can cause an individual electrode's connection with the scalp to be compromised which results in an excessively noisy or completely flat signal for that channel. These are natural products of collecting EEG data.

To correct for these we adjusted the data from each participant using the Brain Vision Analyzer (Version 2.1.1, Brain Products GmbH, Munich, Germany) software suite. We visually inspected the channel streams of participants to identify flat channels or channels with bad signal. These low-quality channels were marked and removed from the dataset, and later reintroduced using interpolation via spherical splines. This process ensures that all participants have similar data shapes for the model to process. To reduce the size of the data we then downsampled the signal to 250Hz from the original 500Hz. We also re-referenced from the original reference electrode to the average mastoid reference for improved resilience to general noise and ran a dual pass phase free Butterworth filter (pass band: 0.1 Hz to 30 Hz; notch filter: 60 Hz) to remove environmental and electrical noise.

We converted the data from EEG stream information to epochs by extracting the -1000 ms to 2000 ms window around each symbol onset event. We used a large time range initially to improve our ability to correct for eye blinks and movement artifacts. The identification for those repetitive artifacts was done using independent

component analysis (ICA) [22], specifically a restricted fast ICA with classic PCA sphering. This process continued until either a convergence bound of 1.0 x 10-7 or 150 steps had been reached. We manually inspected the component head maps and related factor loading to identify ocular artifacts and corrected for these using ICA back transformation.

We then re-segmented the data to epochs with a smaller 1000ms window following stimulus onset, which is the time length used for the actual models. The EEG signal can periodically drift over time, which may make it difficult to compare similar stimuli across exposures, so we performed baseline correction for this using the 200ms prior to the stimulus onset.

While these methods are effective for reducing noise and artifacts in the EEG data, some events may make the data too unusable even after these corrections. For example, if a subject sneezes there is little correction that can be done to improve the signal. Therefore, these cases must be identified and removed from the dataset so they do not confuse the models. This process is called artifact rejection. The artifact rejection utility analyzes every channel on every exposure and removes the exposure if it either contains an absolute difference between the lowest and highest voltage of more than $100\mu V$ on that channel, or if the increase between any two samples on any channel for any exposure was more than $10\mu V/ms$.

3.3 Experiment Methodology

3.3.1 Overview

The experiment methodology follows Sudre et al. [38] using the **2 vs. 2 test**. We train a series of machine learning ridge regression models using the EEG data to generate the individual values for the given indices of **word vectors** matching our word set as the models' predictions.

3.3.2 Word Vectors

We use the **Skip-Gram** word vector set from Mikolov et al. [24]. These word vectors are generated by a neural network with a single hidden layer containing 300 neurons. The neural network is trained to perform a word collocation task: the network receives a single word as input and predicts the probable collocated words for that input word.

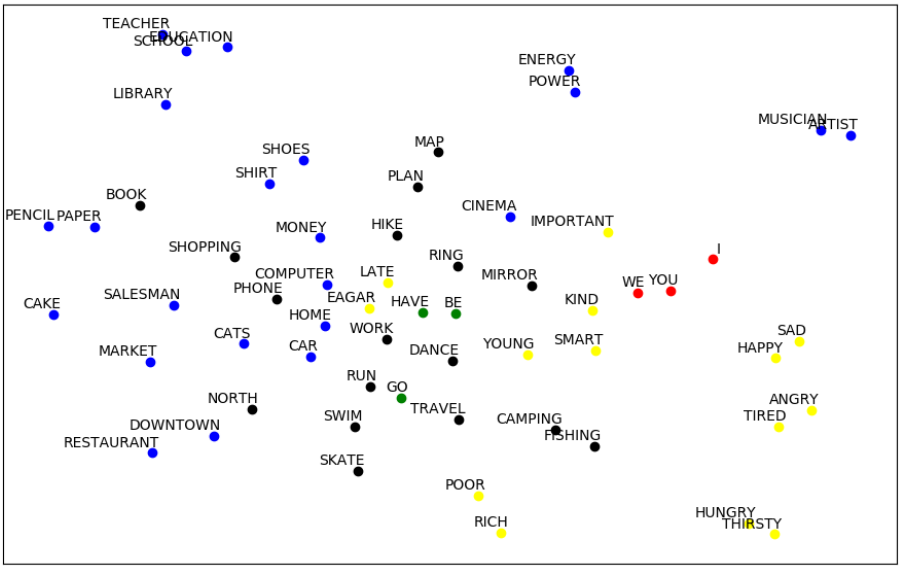


Figure 3.2: A visualization of the word vectors utilized in this thesis, generated by a t-Distributed Stochastic Neighbor Embedding [23]. This technique reduces the input into two dimensions, allowing us to visualize and approximate their relationships in high dimensional space. Similar word semantics are seen clustered closer together. Blue represents abstract and concrete nouns, yellow represents adjectives, red represents pronouns, green represents verbs, and black represents words which act as multiple parts of speech.

Pairs of words are generated from the Google News text corpus using a sliding window over the text corpus. After training, the weights of the model can approximate the probability of collocated words. For each word in the vocabulary, the corresponding weights are extracted from the weight matrix which connects the input layer with the hidden layer. This extraction is done by multiplying the weight matrix with a one-hot input vector representing the target word. The resulting 300-dimensional word vectors are used as training data in our experiment.

Skip-Gram word vectors are a reasonable proxy for word semantics, and have interesting linguistic properties. For example, the difference between the vectors for “man” and “woman” is similar to the distance between the vectors for “king” and “queen”, and the distance between “walked” and “walking” is similar to the distance between “swam” and “swimming”.

More rigorously, Hollis et al. showed that Skip-Gram could predict human judgments for semantic tasks (e.g. sentiment ratings) [13]. Hill et al. additionally concluded that Skip-Gram performs well on their SlimLex-999 evaluation, a high quality word similarity benchmark for computational models of word meaning [12]. Further, Murphy et al. showed that computational models can perform similarly to human benchmarks in the specific context of neurolinguistic decoding tasks [31], and subsequent work showed specifically that Skip-Gram could be used to identify the semantics of many word types in fMRI and EEG [43]. The semantic properties of these word vectors make them a useful tool for performing semantic analysis on brain data. A simplified representation of the word vectors used in this thesis, generated using the t-SNE dimensionality reduction technique [23], is shown in Figure 3.2.

3.3.3 Prediction Model

After the data preprocessing steps mentioned in Section 3.2.1, every participant-symbol pair is represented by a tensor $D \in \mathbb{R}^{(r \times n_e \times l)}$, where r can be between $0 \ldots n_t$, n_t is the maximum number of possible trials seen for a given symbol, n_e is the total number of electrodes, and l is the number of time steps. Due to the randomness of the paradigm, r varies across Ds. Each trial is a matrix in D with dimension $n_e \times l$. Further, we use n_p to denote the number of participants and n_s to denote the number of symbols.

Depending on the type of analysis being performed, we select some trials from the set of all D. The selection process may choose all D for certain participants or

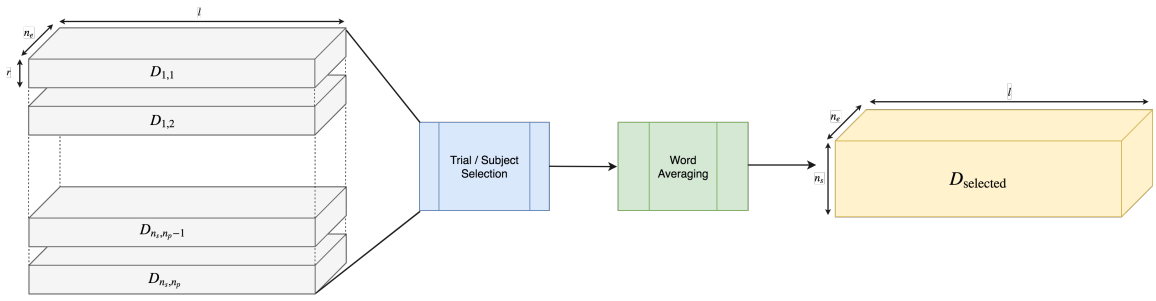


Figure 3.3: The trial selection pipeline. Our initial data contains participant-word pairs D for n_p participants and n_s words that each contain between 0 and n_t trials (r). The trials are of length l and are recorded with n_e electrodes. We select some subsection of these trials, and then average the data across participants to generate D_{selected} which contains the averaged trials for each word.

choose certain trials from each D (see Chapter 5 and Chapter 6 for more details). We average across all participants and trials to create a tensor of dimension $n_s \times n_e \times l$, denoted as D_{selected}. This selection and averaging process is shown in Figure 3.3.

Before we train regression models, D_{selected} is reshaped to produce a matrix with dimensions $X \in \mathbb{R}^{n_s \times (n_e * l)}$. With a sampling rate of 250Hz for a 700ms window with 61 data sensors there will be $61 * 175 = 10675$ numerical features for each sample. The Skip-Gram word vectors also form a matrix with dimensions $Y \in \mathbb{R}^{n_s \times v}$. We find a weight matrix W by training v independent regression models, such that we have one model to predict each dimension of the Skip-Gram word vector set. We use a linear least squares loss function and l2-norm regularization (ridge regression):

$$\underset{W_{:,i}}{min} ||XW_{:,i} - Y_{:,i}||_2^2 + \alpha||W_{:,i}||_2^2 \tag{3.1}$$

where regression model i is trained to predict the ith dimension of the word vectors (column vector $Y_{:,i}$) using weights $W_{:,i}$. The symbol : indexes every element in the dimension, here indicating the selection of a whole row or column vector from a matrix. The notation $||x||_2$ represents the L2-norm of vector x, also known as the Euclidean length. The superscripts represent a traditional exponentiation by two. α is a hyperparameter that controls the level of regularization. We use a standard $\alpha = 0.1$, although we tested several values empirically and found the only minor variation in performance. Using a trained regression model, we can predict a single element of a word vector for a given input $X_{i,:}$ via $\hat{Y}_{i,j} = X_{i,:} \cdot W_{:,j}$.

W is the concatenation of the individual model weights such that $W = [W_{:,1}, W_{:,2}, ..., W_{:,v}]$. Collectively, W is a single model that produces predicted word vectors using $\hat{Y} = X \cdot W$. An example evaluation of the model on a single input vector $X_{i,:}$ from X is seen in Figure 3.4.

A linear model is chosen primarily for consistency with prior literature, but additionally has the benefit of functioning well with small training datasets. Other models, such as neural networks, may be unstable or overfit with small datasets and may also take longer to train as they do not have a closed form solution in all cases.

3.3.4 Evaluation Framework

The set of ridge regression models are then evaluated in a "leave two out" fashion by a binary comparison known as the 2 vs. 2 test. We hold out pairs of symbols $(Y_{i,:}, Y_{j,:})$, and train ridge regression models to predict the vectors of the associated words using

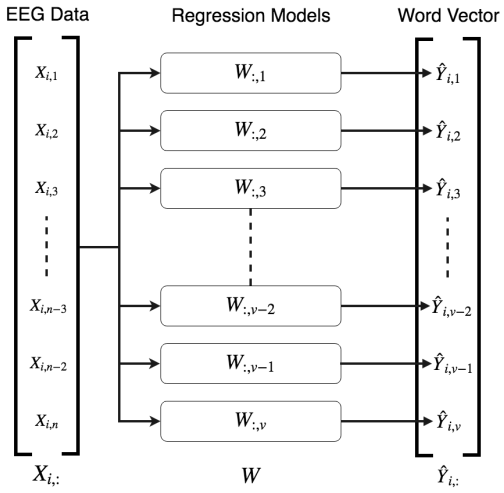


Figure 3.4: An evaluation of the trained model that predicts a word vector from EEG data. The set of regression models can be viewed collectively as a single model that takes a single averaged EEG trial $X_{i,:}$ as input and predicts a word vector $\hat{Y}_{i,:}$ as output. W is the learned weights from the regression models. The number of EEG features is $n = n_e * l$, which varies depending on the experiment, and v which is the length of the word vectors (for our experiments, $v = 300$). Note that the visual representations of the vectors are transposed from their actual shape.

the EEG data from the remaining $n_s - 2$ symbols. The trained model is used to predict the two target word vectors $\hat{Y}_{i,:}$ and $\hat{Y}_{j,:}$ from the held out EEG data. The true word vectors ($Y_{i,:}$, $Y_{j,:}$) are then compared to the predicted word vectors ($\hat{Y}_{i,:}$, $\hat{Y}_{j,:}$) using a vector distance metric d (in our case the cosine distance). The 2 vs. 2 test is considered successful if the sum of the distances between the correctly matched true and predicted word vectors is smaller than the distance of the mismatched vectors as in:

$$d(Y_{i,:}, \hat{Y}_{i,:}) + d(Y_{j,:}, \hat{Y}_{j,:}) < d(Y_{i,:}, \hat{Y}_{j,:}) + d(Y_{j,:}, \hat{Y}_{i,:}) \tag{3.2}$$

We run this test for all possible $\binom{n_s}{2}$ pairs of words. The 2 vs. 2 test can detect if the EEG data is correlated with the word vectors. If the EEG data is not correlated to the word vectors, the 2 vs 2 accuracy (the percentage of the $\binom{n_s}{2}$ 2 vs. 2 tests correct) will be near the chance value of 50%. An example of a 2 vs. 2 comparison is shown in Figure 3.5.

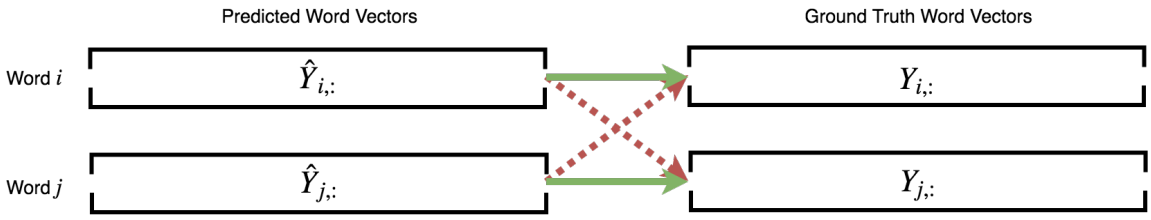


Figure 3.5: A 2 vs. 2 comparison. We perform this comparison for all possible pairs of words, and the resulting 2 vs. 2 accuracy is the percentage of total symbols which are correctly aligned (when measured by a standard vector distance metric).

3.3.5 Validating Statistical Significance

We tested the statistical significance of our results from experiments in Chapter 4 using permutation tests. For each experiment, we reran the pipeline, but randomly shuffled the order of the word vectors so that the true word vectors no longer correctly matched with the EEG data for each symbol. This randomization was done after averaging over participants and symbols. We ran the same experiments on 300 permutations of the data, and used the resulting 300 2 vs. 2 accuracies to approximate the null distribution (where the data and labels have no statistical relationship).

As expected, we found that the empirical null distribution had a mean close to 50% (chance accuracy) for all experiments. The p-values were obtained by testing the reported accuracy against a Gaussian kernel density estimation fit to the associated empirical null distribution. We corrected for multiple testing using the Benjamini-Hochberg-Yekutieli procedure where applicable [3], with an alpha value of 0.05.

3.4 Conclusion

In this section we introduced the collection experiment, in which participants performed a reinforcement learning task designed to emulate language learning. With data from these 25 participants we performed standard EEG preprocessing techniques. The preprocessed data from these participants is provided to a model which correlates the EEG data with semantic word vectors by building a set of machine learning models that can be used to predict word vectors for unseen words.

We also introduced the 2 vs. 2 test, which is used to test and statistically validate whether or not the model could learn a mapping from the data. In the following chapter we will introduce the experiments that we perform on the model so we could learn more about how the brain processes language.

Chapter 4

Experiment Designs

We devised six experiments to study the emergence of semantic representations in the brain during the symbol learning paradigm. We discuss these in order, beginning with the more fundamental experiments and ending with more detailed analyses built on those earlier experiments.

4.1 Semantic Representation Experiment

When the participants initially viewed the symbols, they did not know the English equivalent. But, as they learned the symbol mapping through trial and error, semantic understanding developed. Our methodology allows us to test if word semantics are correlated to EEG activity while viewing the corresponding symbol, and this first experiment is the simplest test of the methodology. We hypothesized that we would see statistically significant accuracy when training on trials after learning is assumed to have occurred.

For each participant, we removed symbols that were presented less than six times to that participant, and removed the first two exposures of each symbol for each participant (as we can assume the symbols were not associated with a semantic meaning in the first two trials). The parameters for trial selection (two and six) provide a balance between allowing the participant time to learn the word, while also retaining enough data to train our model. We expected longer trial lengths to have higher accuracy, as it includes more time for the participants to think about the translation, but we also do not want to include muscle artifacts. Therefore, we cut trials to 700 ms in length to avoid including muscle signals in the exposure after the appearance

of word choices at 500 ms (as the shortest response time was 248 ms). We also test with trials cut to 500 ms to compare the 2 vs. 2 accuracy when the appearance of word choices is excluded from the exposure, to verify that we are not inadvertently identifying the semantics of the displayed words. We then averaged the remaining exposures over all participants for each word, which gave us a single noise-reduced exposure per word. This data was used to train regression models and perform the 2 vs. 2 test.

This experiment is the closest to previous work in other modalities, and functions as our baseline validation of the experiment methodologies discussed in the previous chapter. The data extracted is intended to be representative of words the participants have learned the mapping for, although it will not be identical to the task of reading English words as the participants must still perform the cognitive task of correctly identifying the mapping for the symbol. A statistically significant 2 vs. 2 accuracy in this experiment indicates that we can correctly identify the semantic representations of the mapped English words in EEG.

4.2 Participant Learning Experiment

The previous experiment allowed us to detect if semantic representation of learned symbols could be detected using EEG. To build on this, we can leverage the unique nature of this artificial language paradigm to better understand how semantic representations develop as participants learn a language mapping. To do this we tested *when* we can detect the semantic mapping, as a function of the number of exposures. Here we determined if we can detect the average onset of symbol learning. We compared the 2 vs. 2 accuracy for the earlier trials (e.g. trials 1-3, before the symbol meaning was learned) to the later trials (e.g. trials 4-6, after participant learning) to test if we can measure the emergence of semantics during the paradigm. As in Section 4.1, we only considered participant-word pairs with six or more exposures, to ensure a fair balance in the number of exposures being included in each group. We also cut the trials to 700 ms and 500 ms as before and followed a similar averaging strategy. We compared the 2 vs. 2 accuracy of averaged overlapping subsets of three exposures, selected from the first six exposures.

We hypothesized that the 2 vs. 2 accuracy would increase in later exposures, as the symbol mapping was learned by the participants in the reinforcement learning paradigm. It is important to reiterate that we are detecting the semantics of the

English word, not of a representation for the symbol itself. Therefore, accuracy will only increase if participants are able to successfully think of the English translation for the symbol. While they were given no specific instruction to think of or visualize the English concept, we hypothesized they will do this intuitively as a requirement for responding in the task.

4.3 Reward Positivity Experiment

We also wanted to quantify learning using more traditional learning measurement mechanisms. Typically, this measurement is done by comparing the amplitude of the reward positivity over trials [42]. We expected that with this experimental paradigm we would see reward positivity for the earlier trials, and diminish thereafter. Our application of 2 vs. 2 accuracy to measure learning is novel. This more standard analysis is meant to provide evidence that participant learning could be detected using the EEG data. We hypothesized that both this experiment and the Participant Learning Experiment (Section 4.2) would show the effects of participant learning.

It is important to reiterate here that the Reward Positivity Experiment was performed by Chad C. Williams. This section is included in this thesis for a contextual comparison with the results found by our new approach.

This experiment consisted of two parts. Firstly, we divided the trials into groups of correct and incorrect responses then averaged across all participants and symbols. We compared the amplitudes of the average signals at the FCz electrode, where the reward response is the strongest. Secondly, we compared the amplitude of the first six correct responses (averaged in a similar fashion) at the same FCz electrode. Note that these six responses *cannot* be directly compared to the six responses in the Participant Learning Experiment (Section 4.2) because the Participant Learning Experiment considers all trials, whereas the present analysis considers *only correct trials*. To determine 1) the amplitude of the reward positivity and 2) the change in correct waveforms as learning progresses (first six correct trials), a max peak time was first extracted from the reward positivity difference waveform for each participant. An averaged max peak at 278ms was found within the 250 ms - 400 ms time range. To extract the amplitude of the reward positivity and correct waveforms, we averaged the data +/- 25 ms surrounding this peak.

4.4 Participant Performance Experiment

Reward positivity correlates to participant learning as measured by behavioral feedback, so we also validate our measurement using participant responses during the paradigm. Recall that we recorded the participants' behavioral responses as they learned to map the symbols to English words. Participants with higher behavioral accuracy learn the symbol mapping faster, and should therefore have a stronger representation of the associated word semantics. As in the the Reward Positivity Experiment (Section 4.3), this could provide evidence that we are able to detect participant learning, and even quantify the efficacy of learning. We hypothesized that the behavioral accuracy of the participants should be correlated to the average 2 vs. 2 accuracy for participants grouped by behavioral accuracy.

To combat the noise inherent in EEG, the 2 vs. 2 accuracy was calculated over the average of several participants. Here we considered two groups: those 7 participants with the highest and the lowest behavioral accuracies (these groups were chosen by the natural grouping in their task accuracies). We cut the trials to 700 ms and 500 ms in length and averaged across trials as in the prior two experiments. We then calculated 2 vs. 2 accuracy over these two groups, and compared the groups' average task accuracies. We hypothesized the 2 vs. 2 accuracy and behavioral accuracy should be positively correlated.

While this experiment is designed to provide some insight into the relationship with task accuracy, segregating the participants into separate groups affects the ability of the averaging process to combat noise. This has a progressively negative effect on 2 vs. 2 accuracy, which is important to consider when evaluating conclusions from this experiment.

4.5 Time Windowing Experiment

We also wished to understand the recollection of a semantic representation when evoked by a newly learned symbol. Here, we could take advantage of EEG's high temporal resolution to analyze the brain's processing of symbols over time. We did this by separating the averaged EEG data into time windows, each 50ms long, and then evaluating the model pipeline on only the EEG data within a window. This is an additional filtering step on D_{selected} that reduces the dimensions of D_{selected} to $\mathbb{R}^{n_s \times (n_e * s_l)}$ where the length of the selected time window is defined as $s_l \leq l$. This

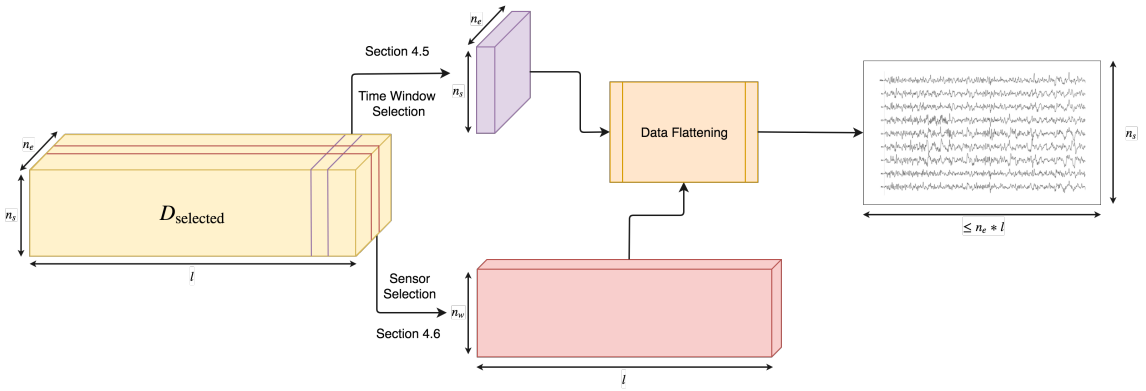


Figure 4.1: The data reshaping pipeline. Averaged trial data from D_{selected} can be directly passed to the reshaping process, or it can be passed through another selection. We can perform two types of selection, one for time analysis or one for channel analysis. We reshaped to flatten across the electrode dimension, such that our training data for the model is $n_s \times (n_e * l)$.

process is shown in Figure 4.1 as the "Time Window Selection" step. The 2 vs. 2 accuracies from these groups can be visualized as a graph over time.

The time window with the highest accuracy represents the peak in time that the semantics are most strongly represented in the brain data. In their MEG experiments, Sudre et al. [38] found that the decodability of nouns peaks in the 350ms-400ms time period after stimulus onset. We hypothesized that the peak accuracy for our experiment would be later, as participants must map each symbol to the English counterpart.

4.6 Sensor Selection Experiment

In addition to the timing of semantic representations, we were also interested in the *localization* of semantic representations. Thus, we explored the 2 vs. 2 accuracy when only a subset of electrodes were used as input to the regression model.

This is similar to the Time Windowing Experiment, except here we select only subset of the electrodes and on which to run the model pipeline. This is an additional filtering step on D_{selected}, which reduces the dimensions of D_{selected} such that $D_{\text{selected}} \in \mathbb{R}^{n_s \times (s_e * l)}$ where the number of selected electrodes is defined as $s_e \leq n_e$.

This process is shown in Figure 4.1 as the "Sensor Selection" step. We created n_s groups of sensors, where each sensor has its own group that consists of itself and its immediately neighboring sensors. This allowed us to explore individual sensor

accuracies while being less sensitive to that individual sensor's noise. Each sensor has between one and four neighbors under our mapping. We can visualize these results using a topographic plot, where the value of each electrode is the 2 vs. 2 accuracy of the corresponding group in which that electrode is the main electrode. We ran this analysis for three time windows to better understand the localization in separate time periods: 0 - 500 ms, 500 - 1000 ms, and 0 - 1000 ms.

4.7 Validating Statistical Significance

We tested the statistical significance of our results from experiments in Chapter 4 using permutation tests. For each experiment, we reran the pipeline, but randomly shuffled the order of the word vectors so that the true word vectors no longer correctly matched with the EEG data for each symbol. This randomization was done after averaging over participants and symbols. We ran the same experiments on 300 permutations of the data, and used the resulting 300 2 vs. 2 accuracies to approximate the null distribution (where the data and labels have no statistical relationship).

As expected, we found that the empirical null distributions across all experiments had an average mean of 50.01% (chance accuracy) with $\sigma = 0.4\%$. The p-values were obtained by testing the reported accuracy against a Gaussian kernel density estimation fit to the associated empirical null distribution. We corrected for multiple testing using the Benjamini-Hochberg-Yekutieli procedure where applicable [3], with an alpha value of 0.05.

We also utilized nonparametric bootstrap testing, which allowed us to compute a statistical significance score for the difference in 2 vs. 2 accuracies across two experiments [9]. To perform bootstrapping, we sampled with replacement n_p times from the list of participants and reran the model pipeline on those sampled participants with the parameters of each experiment. This is repeated R times. Similar to the permutation test, the resulting 2 vs. 2 accuracies form empirical distributions for each experiment. We generate a normal-theory confidence interval around the real 2 vs. 2 scores using the respective empirical distribution. We compare two of these confidence intervals, one for each experiment, to test if two 2 vs. 2 scores for an experiment are statistically different. In this work here we use $R = 100$ and generate the confidence intervals with $p < 0.05$.

4.8 Conclusion

In this section we introduced the six experiments that we perform on the collected dataset. Our first experiment performs a baseline test to see if we can detect semantics in EEG using a similar methodology which has been applied to fMRI and MEG data. The second experiment expands on this to measure how the semantic representations develop over the course of the participants' learning. The third experiment performs a traditional reward positivity analysis, so we can compare the behaviour between the two analyses. The last two experiments break down the analysis in terms of time and sensors, to help us understand when the semantic representations peak and from what areas of the brain.

With our collection paradigm described, the experiment framework in place, and all of the experiments detailed, the next chapter will discuss the results found in these experiments.

Chapter 5

Experiment Results

In this chapter we will discuss the results of the experiments that we performed on the reinforcement learning dataset. We will discuss these in the same order that they are described in the previous chapter, Chapter 4. Lastly, we will conclude with a summary of our findings. The following chapter includes discussion on the conclusions we can draw, and the new discoveries we found.

5.1 Semantic Representation Experiment

This first experiment, the Semantic Representation Experiment, was designed to determine whether we could identify semantics in EEG under this reinforcement learning paradigm. We utilize a similar approach that has detected semantic representations successfully in fMRI and MEG. Here, we trained our model on a subset of the trials where we anticipated participant learning would have occurred, and evaluated this model with the 2 vs 2 test.

Our model produced a 2 vs 2 accuracy of **79.54%** for the 0 - 700 ms window which is statistically above chance with $p < 0.001$. For the 0 - 500 ms window, the 2 vs. 2 accuracy is **69.15%** which is also statistically above chance with $p < 0.001$. This shows we can detect semantic representations in the brain using EEG data. While applying this experiment in this paradigm and with this collection methodology has some unique attributes that will be further discussed in Section 6.1, the main value of this experiment is that it provides a baseline validation for using this methodology to explore semantic representations with EEG in more detail in later experiments.

5.2 Participant Learning Experiment

This experiment builds on the Semantic Representation Experiment, and aims to identify the trial at which we can detect meaning. To do this we utilize a sliding window with a window size of three trials over our included dataset. Figure 5.1 plots the 2 vs. 2 accuracy over trials using this sliding window technique for the 0 - 700 ms time period. When we average exposures (1, 2, 3) of each symbol, we achieve a 2 vs. 2 accuracy of **46.70%**. Exposures (4, 5, 6) produce an accuracy of **72.35%** with $p < 0.001$ (FDR corrected). We see a similar pattern in the 0 - 500 ms time period with accuracies of 55.87%, 56.27%, 58.53%, and 64.86% for each sliding window respectively. We applied bootstrapping to generate normal theory confidence intervals for both the first and last sliding window, which confirmed with $p < 0.05$ that there is a statistically significant difference in 2 vs. 2 accuracy between the first trials participants see (1, 2, 3) and the later trials participants see (4, 5, 6).

Due to a reduction in data, the 2 vs. 2 accuracy over trials (4, 5, 6) is slightly lower than the 2 vs. 2 accuracy in the previous experiment, which used trials beyond

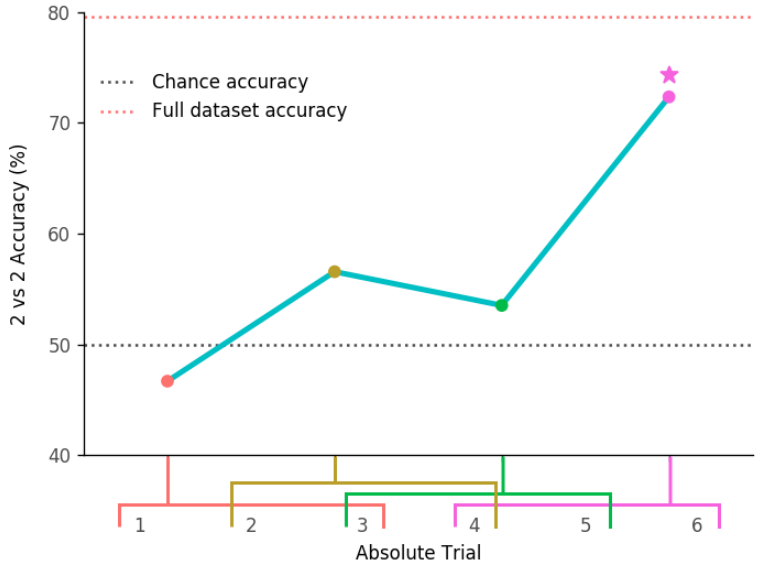


Figure 5.1: A graph of 2 vs 2 accuracy over trials in the 0 - 700 ms time period. This graphic shows how participant learning develops over time with a sliding window of three trial averages. 2 vs 2 accuracy increases notably in the last window, showing learning occurs in the later half of trials. A star indicates a statistically significant value with $p < 0.001$ (FDR corrected).

the 6th exposure. However, this result confirms we can detect participant learning with our model. This is a novel and unique way to analyze the process of learning an artificial language mapping, which has benefits that will be further discussed in Section 6.2.

5.3 Reward Positivity Experiment

In order to compare our learning analysis model (described in the previous section) with a more traditional ERP based analysis, here we analyze the reward positivity in the FCz electrode. We perform a comparison of the correct and incorrect responses as well as a comparison of the first six correct responses for a word to do this.

It is important to reiterate here that the Reward Positivity Experiment was performed by Chad C. Williams. This section is included in this thesis for a contextual comparison with the results found by our new approach.

Figure 5.2 shows the presence of reward positivity, confirmed by a dependent samples t-test of the difference waveform between the first correct response and the average of incorrect responses with $p < 0.001$. Figure 5.3 shows the individual correct trials averaged over participants. Here, we see a strong reward positivity for the first correct trial and a diminishing effect on subsequent correct trials (fitting a power law function with $R^2 = 0.96$). This confirms our hypothesis that the reward positivity decreases and the 2 vs. 2 accuracy increases over trials.

5.4 Participant Performance Experiment

Here we test if the participants' 2 vs. 2 accuracies are related to the participants' average task accuracies by examining the behavioral data. The average task accuracy of individual participants ranges from 72% - 90% and the mean over all participants is 81%, and the standard deviation is 4%. While the average task accuracy for participants may not be completely comparable across participants, as it may include different number of total trials for each participant, it still functions as a representative number of the general rate at which the participant learned the mapping. Participants who learned the mapping faster would have higher task accuracy in each block and a higher average task accuracy, while participants who learned the mapping slower would have to repeat more blocks and have their average task accuracy reduced by the poor performance on those blocks.

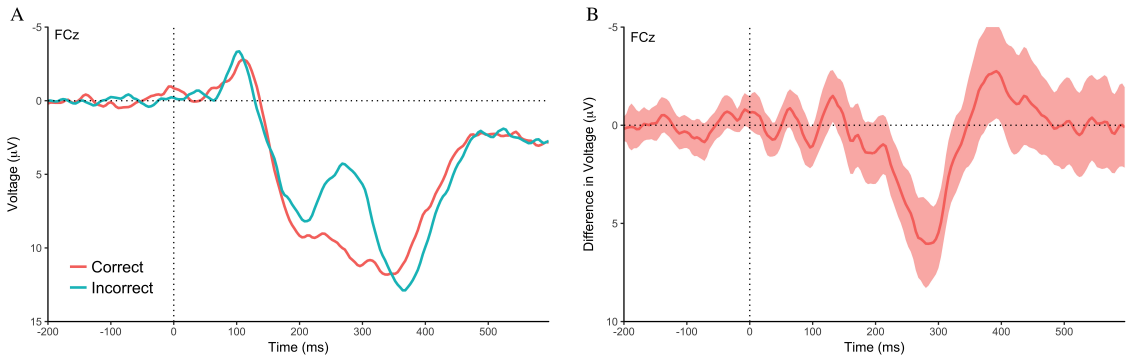


Figure 5.2: The reward positivity for correct and incorrect responses at the FCz. In both graphics, the y-axis is positive downward. In **A**, we see the signals of the averaged first correct trials for each word and all averaged incorrect trials. In **B**, we see the difference between the averaged first correct trials and all averaged incorrect trials with 95% confidence intervals. There is a clear presence of the reward positivity. This figure courtesy of Chad C. Williams.

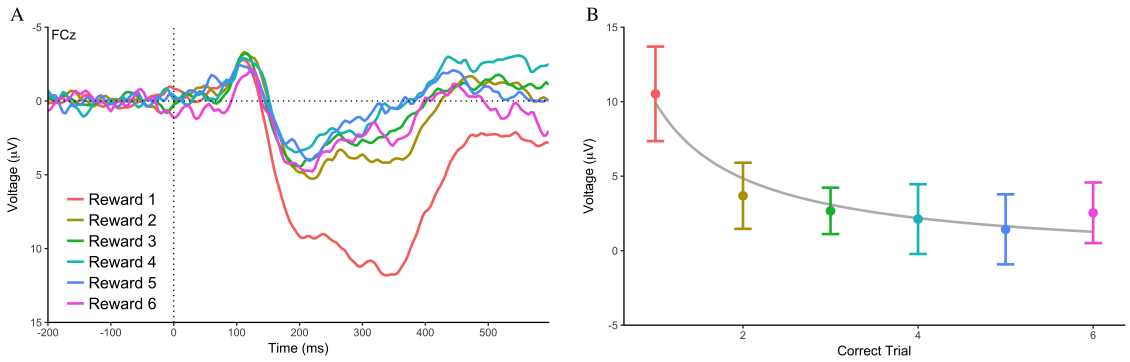


Figure 5.3: The reward positivity between the first six correct responses. The y-axis is positive downward for the left subfigure and positive upward for the right subfigure. In **A**, we see the amplitude of the signal at the FCz electrode between the first six averaged correct responses. The amplitude of the correct waveform of the reward positivity is large on the first correct trial and diminishes with subsequent rewards. The change in this waveform indicates a diminishing reward positivity across learning. In **B** we see the amplitude during the highest period of 228ms - 328ms after stimulus onset compared between the first six averaged correct responses. Again, here we see a clear reward positivity in the first correct trial and a diminishing effect on subsequent correct trials. This figure courtesy of Chad C. Williams.

We split the participants into two groups based on their task accuracy, and evaluated the 2 vs. 2 accuracy within these two groups. Because the variance of average task accuracy is small across participants, we evaluated small groups of top and bottom performers. The average 2 vs. 2 accuracy in the 0 - 700 ms time period of the 7 participants with task accuracy below 80% is **59.71%**, and the 2 vs. 2 accuracy of the 7 top participants (all above 85% task accuracy) was **65.13%**. While both of these 2 vs. 2 accuracies are significantly lowered due to the reduction in training data compared to the previous two experiments, this suggests a relationship between task performance and our ability to detect the semantic meaning of the symbols via EEG. However, this effect is less obvious in the 0 - 500 ms time period in which we see bottom and top accuracies of 57.47% and 57.55%, respectively.

5.5 Time Windowing Experiment

The Time Windowing Experiment allows us to examine when the semantic representation in the brain is the strongest. The 2 vs. 2 accuracy as a function of time *within* an exposure is shown in Figure 5.4, allowing us to pinpoint the window where accuracy peaks. We find accuracy peaks in the 600ms-650ms window, at 74.57% ($p < 0.001$, FDR corrected). We also see an earlier spike which peaks in the 150ms-200ms window, at 74.34% ($p < 0.001$, FDR corrected).

The later peak confirms our hypothesis, which was that we would see a delayed peak due to the cognitive requirement of translating from the symbol to the English word before the semantics of the English word can be represented. However, we were surprised to see a strong early peak in the semantic representation. There is some evidence of an early semantic representation signal in other work, and the later effect may be conflated with the appearance of word choices at 500ms, which is discussed in more detail in Section 6.5.

5.6 Sensor Selection Experiment

In this experiment we test which areas of the brain are contributing the most to the 2 vs. 2 accuracy. We categorized sensors into n_s groups for analysis, where each group consists of a primary sensor and its neighboring sensors. We used the accuracy of each sensor group to annotate the accuracy of the primary sensor, and then performed a topographic interpolation of the 2 vs. 2 accuracy over the brain.

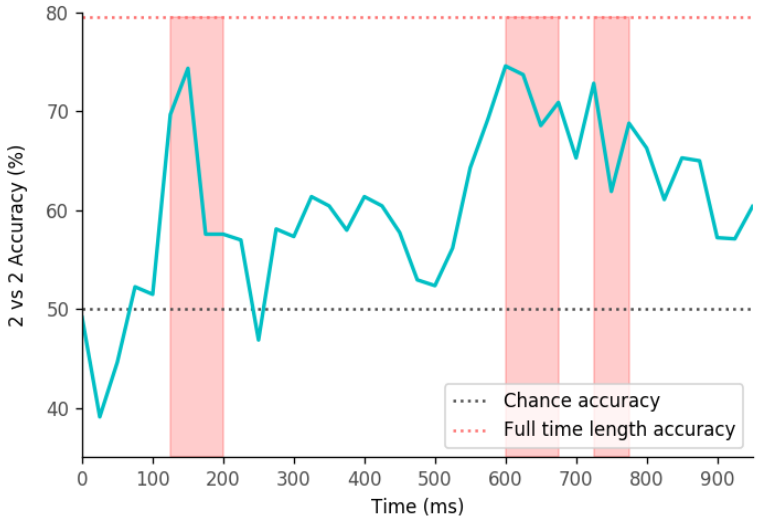


Figure 5.4: A graph of 2 vs 2 accuracy over time. This graphic shows scores on the 2 vs 2 test evaluated with only the data present in 50ms incremental windows. For example, the point at 25ms defines the 2 vs. 2 accuracy over the 25ms-75ms period. Statistically significant time windows are identified in red. The highest performing period is 600-650ms with 74.57% accuracy ($p < 0.001$, FDR corrected). We also see an earlier spike which peaks at 150ms-200ms with 74.3% accuracy ($p < 0.001$, FDR corrected).

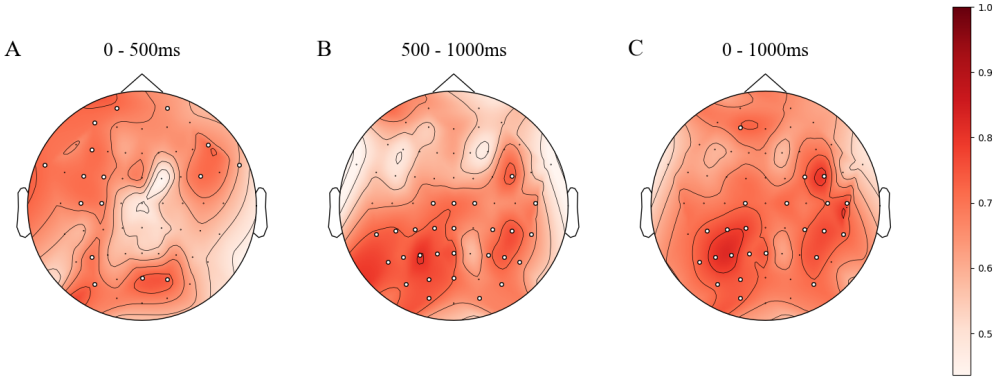


Figure 5.5: The results of the model on various brain regions. Each sensor represents a group containing it and its immediate neighbors, and we calculated 2 vs. 2 accuracy using single groups. A topographic plot of the 2 vs. 2 accuracies is shown for three time periods: **A** 0–500ms window; **B** 0ms–1000ms window; and **C** 0–1000ms window. Statistically significant groups are shown in white ($p < 0.001$, FDR corrected).

The topographic interpolation of three time windows is visible in Figure 5.5. We see lower 2 vs. 2 accuracies for the earlier time window (0-500ms) compared to the later time window (500ms-1000ms), however we see the highest accuracies over the entire time window (0-1000ms). Removing a substantial amount of data effects accuracy negatively, however a large amount of the sensors remain statistically significance, which indicates that the semantic representation is highly distributed across the cortex.

5.7 Conclusion

In this chapter we have described the results found in our experiments. The next chapter will explore these in more detail, including the conclusions we can draw and an investigation of how our work fits into alignment with existing literature. Our early high level analysis of these results shows that it is possible to detect the semantic representation of the English words in EEG and that we can detect how this semantic representation emerges during learning over time. We can also see evidence that our results align with other measurements of learning, such as the reward positivity and task accuracy. Lastly, when exploring the 2 vs. 2 accuracy over both time and sensors we see interesting spikes both early and late and a distributed source of semantics across the human cortex.

Chapter 6

Discussion and Analysis

We have compiled evidence that we can detect semantics and learning using EEG with statistically significant accuracy. While previous work has shown that learning can be detected via EEG, our work is the first to show that the product of that learning (i.e. the semantic representation) can be detected via our machine learning methodology, and that it aligns well with more traditional ERP-style analyses. Here in this chapter we will explore our results in more detail, along with the conclusions we can draw from it.

6.1 Detection of Semantic Representation

We were able to achieve a 2 vs. 2 accuracy of 79.54% in the Semantic Representation Experiment, which provides evidence that there is a strong correlation between the EEG data of participants and the word vectors mapped to each symbol. This confirms our hypothesis that we would see statistically significant accuracy on the 2 vs. 2 test. Previous work has mostly focused on MEG and fMRI [27, 38], but our results show that a similar word decoding methodology can also be applied to EEG data even when the paradigm differs.

The accuracy is higher for the 0 - 700 ms window and lower for the 0 - 500 ms window, but both are statistically above chance. This makes intuitive sense, as there is less data for the model to leverage in the 0 - 500 ms window. Additionally, at 500 ms the appearance of word choices occurs. This reduces the possible translations to only four words, which aids the participants and likely provides valuable decoding information for the model.

Murphy et al. also showed this methodology has promise in EEG [29], but was focused on two semantic categories of concrete nouns (tools and mammals) while we expand to a much more varied vocabulary and utilize a different paradigm based on reinforcement learning. Though EEG has its limitations, it is sometimes preferable due to reduced cost and improved portability over MEG and fMRI. This makes the adaption of this methodology to EEG an important contribution and goal for research in this area, as it lowers the costs for semantic representation research by several orders of magnitude. Additionally, there are individual benefits compared to each modality such as the increased time resolution of EEG compared to fMRI or increased sensitivity to more brain areas compared to MEG.

It is also notable that our stimuli were symbols rather than wordforms. Previous studies used written words alongside illustrations [27, 38], and recent work showed that accuracy is higher when utilizing illustrations for word context [33]. Critiques of earlier work posited that the models were simply leveraging the brain's visual representations, such as the dominant shapes in the image or the length of the word, and might not be related to word meaning. Because the mapping of symbols to words was totally arbitrary, our results are strong evidence that models leveraging word vectors truly model the brain's representation of word *meaning*, and not some low-level visual features of the stimuli. Although the translated word did become visible as one of the four options at the 500 ms mark, we see statistically significant 2 vs. 2 accuracy even when only training on the period that the symbol alone is visible.

In previous work, participants were requested to visualize the concepts while viewing words and images [27, 38]. In our experiment, participants were provided no instructions regarding visualization, only instructed to perform the reinforcement learning task. This provides evidence that the semantic representations are detectable in a more complicated word mapping task and not only when performing the simple visualization tasks. This also shows that participants do not need to be coached to explicitly visualize the concepts in order to detect semantics.

We also achieved high accuracy on the 2 vs. 2 test using a more direct and simple averaging mechanism. While some previous literature has approached this by grouping and averaging only the exposures of a single participant (giving an averaged result for each participant-word pair) [27, 38], we found that the models performed best when we averaged the sensor readings across all participants, with no extra accounting for the differing shape or size of participants' heads. The smoothness of EEG likely contributes to the success of this approach.

Our work also shows that the word vector approach generalizes to different parts of speech. Many previous studies used only concrete nouns [27, 38, 29]. Although the majority of our words were nouns, participants saw concrete nouns, abstract nouns, pronouns, adjectives, and verbs. Even concrete and abstract nouns, while both the same part of speech, can have different electrophysiological attributes which make semantic modelling a more complicated task [1]. This aligns with a recent study by Pereira et al. which had a similar distribution of parts of speech for an fMRI model but utilized much more training data [33]. To score high on the 2 vs. 2 test the model must be able to discriminate effectively between these categories to some degree. Thus, our methodology can detect a semantic difference between words both within and between various parts of speech.

While this methodology has been applied to other brain imaging modalities before, and we are not the first to show evidence of semantic representation using EEG, this initial experiment contributes to the body of evidence supporting the use of EEG for semantic representation research and provides iteration and new insight on the technique of using linear models to predict word vectors as a function of brain data.

6.2 Measurement of Participant Learning

One of the exciting and unique attributes of this experiment paradigm compared to other experiments in semantic representation research was the introduction of an artificial language through a learning task. This allows us to research and understand how the semantic representation develops in the brain during participants' learning. We therefore hypothesized that the 2 vs. 2 accuracy would increase in later exposures for a given word as participants learn.

Our Participant Learning Experiment measured the strength of semantic representation at different exposure counts for individual words, and found that semantic representation of a symbol emerges after or near the 4th trial. The average of trials (4, 5, 6) produced 2 vs 2 accuracy of 72.35%, which is significantly above chance with $p < 0.001$ (FDR corrected). This is evidence that the symbol mapping is learned very quickly, within only a few exposures. We see this effect in both the 0 - 700 ms time period and 0 - 500 ms time period.

Previous work has detected participant learning by grouping trials together and measuring the reward positivity [17]. However, reward positivity does not always coincide with learning-related behavioral changes (i.e. task accuracy), making it po-

tentially unclear if reward positivity is related to a direct brain function for learning or if reward positivity is an indirect effect from a brain function related to feedback [40]. Recent work has made new arguments supporting the view that reward positivity is a direct index of neural learning [42].

Our work provides another angle for comparison, as we measured learning by detecting the actual *concept* to be learned (i.e. the word vector) rather than using an indirect measurement via an ERP. This supports the claim that reward positivity is directly related to a brain function for learning, or at least indicates evidence for a negative correlation with 2 vs. 2 accuracy. That is: the reward positivity response decreases in amplitude as the representation for the concepts being learned increases in 2 vs. 2 accuracy.

In addition to providing support for the reward positivity, our model has benefits over the traditional analysis method in many paradigms. Since our model detects the semantic representation, it can be used to model both the process of learning and the later retention of learning. Reward positivity can show when learning has occurred, but the absence of reward positivity does not indicate the knowledge of the topic. Thus, our approach could offer benefits in experiments where it is important to measure *retention* of the mapping.

A further benefit of our approach is assessment of the speed of learning, in addition to detecting learning processes. Previous literature in learning detected only the presence of learning, but did not aim to quantify the speed at which learning occurs [17]. Our work shows learning as it occurs over the averaged trials, even in a fairly complicated experimental paradigm. This can additionally be more useful than using task accuracy alone to measure learning, as task accuracy cannot detect guessing. Future experiments may find new insight when researching learning by leveraging the more complete picture provided by considering 2 vs. 2 accuracy.

6.3 Comparison with Reward Positivity

In the Reward Positivity Experiment we compared the amplitude of the reward positivity ERP between both the correct and incorrect responses as well as the amplitude of the first six correct responses. This is not a direct comparison to the Participant Learning Experiment, which includes both incorrect and correct responses in the six trials considered. In this ERP based experiment, we considered only trials where the subject chose correct response. As anticipated, our results showed a measurable

difference in the amplitude of the ERP between the first correct and average incorrect responses. Additionally, we saw a strong response to the first correct feedback and a diminishing response to subsequent correct feedbacks.

These results align with existing studies of the reward positivity in reinforcement learning paradigms, where reward positivity is strong in the first correct trial and then diminishes [17, 18, 2]. They also align with the results seen in our Participant Learning Experiment, as both measurements detect that learning is occurring. Though we cannot directly compare, this provides evidence that our measure of learning (2 vs. 2 accuracy) responds similarly to the traditional method of measuring learning through the reward positivity response.

6.4 Comparison with Participant Performance

In the Participant Performance Experiment, we hypothesized that the task accuracy of participants should be related to the 2 vs. 2 accuracy. To test this, we took both the top 7 and bottom 7 performing participants (measured in terms of task accuracy) and evaluated the model pipeline with those groups in the 0 - 700 ms time period. We found the 2 vs. 2 accuracy is higher (65.13% vs 59.71%) for the group that performed better in terms of task accuracy. Higher task accuracy implies a participant learned the symbol mapping better than a participant with lower task accuracy, and the 2 vs. 2 results show that this trend is measurable in the EEG data. However, we did not find an effect of similar degree for the 0 - 500 ms time period.

The experiment paradigm was designed to ensure that participants had learned the mapping before progressing through the experiment. Because of this design, the average task accuracy of all participants is very close together and this type of effect may be difficult to measure. For these reasons we do not believe we can confidently confirm the experiment's hypothesis here, and additional experimentation is required to more rigorously test and validate this theory.

6.5 Time Windowing Analysis

The Time Windowing Experiment separated the averaged trial data into 50ms windows and evaluated the model pipeline on each window. We had hypothesized that our time analysis would see a delayed peak response compared to prior work [38] due to the additional processing time required to map from the symbol language to the

English word. We found a peak accuracy in the 600-650ms window with a 2 vs 2 accuracy of 74.57%. Previously, a similar experiment using MEG data that did not utilize a symbol mapping component found a peak accuracy of 350ms-400ms [38]. This delay appears to be present in our results, and slows the emergence of semantic representations by about 250ms. This provides evidence that the newly formed mapping delays retrieval by about a quarter of a second.

Interestingly, very little 2 vs. 2 accuracy is lost when we evaluated our model using only a small window (50ms) compared to our initial experiment (Section 6.1) which used a much longer window (700ms). By windowing the exposure we remove the majority of data provided to the regression models, but see only a minimal reduction in accuracy (5%). This is notable since a model with significantly fewer parameters is much faster to train and evaluate.

We also see a large, statistically significant spike in 2 vs. 2 accuracy early in time, around 125ms-200ms. Sudre et al. had also found statistically significant accuracy early in the processing pipeline, but had determined this to be heavily correlated to the visual features of the words they presented rather than the semantics of the words [38]. Our experiment, however, should not experience the same correlation with visual features such as *Word Length* or *Image Diagonalness* as we use an arbitrary mapping of words to symbols (discussed in Section 6.1). It is notable that studies have shown evidence of semantics much earlier in time, such as Moseley et al. who were able to differentiate between semantic categories as early as 150ms into the exposure [28]. There are several key difference between this experiment and Sudre et al.. First, our experiment uses Skipgram instead of vectors based on behavioral responses to semantic questions. The Skipgram method, based on word co-location, encodes semantics differently than the behavioral response vectors, and may model some component of semantic representations that is available earlier in time. Second, our experiment uses multiple parts of speech, whereas the original Sudre et al. used only concrete nouns. Perhaps part of speech identity is differentiable early in time, which would not have been detectable in an experiment using only nouns. Third, the task performed by our participants is much more interactive and requires engagement learning. The Sudre et al. task was a simple semantic question answering task. Perhaps a more engaging task evokes a faster or more consistent neural response. Each of these differences may be the reason we see such early semantic activation. We will need further experiments to determine if this early semantic signal is replicable, and if it is truly correlated to very early semantic recall.

The display of the four word options at 500ms is likely to have an effect on the timeline of these results as well. At the initial display, participants attempted to translate the symbol to an unknown word. With the addition of the four options at 500ms, participants had a narrowed down option set of only four words to map to. This may lead to the appearance of two accuracy spikes: one of early semantics shortly after the initial stimuli and one of confirmation after the four word appearance at 500ms. We can still confirm that the mapping is being measured in both cases, rather than the word, due to the differences in 2 vs. 2 accuracy between earlier and later trials of the same word as seen in the Participant Learning experiment discussed in Section 5.2.

6.6 Sensor Selection Analysis

In the Sensor Selection Experiment, we divided the EEG sensors into groups, and evaluated the 2 vs. 2 accuracy using only the sensors in each group. We visualized the 2 vs. 2 accuracy with a topographic plot, and highlighted the primary electrodes for groups that had significant accuracy. The number of electrodes in each group differ depending on the number of neighboring electrodes, and thus we have varying numbers of features for each 2 vs. 2 experiment. Additionally, the diffusion of EEG signals through the skull and scalp, as well as the overlap between neighboring sensors, mean that each region is not completely independent in the 2 vs. 2 test. While it is important to take these two factors into consideration when interpreting the accuracies, we will explain how the results align with existing literature in other methodologies and our other experiments.

We see lower 2 vs. 2 accuracies for the earlier time period (0–500ms) and higher accuracies for the later time period (500ms-1000ms). However, we see the highest accuracies when using the whole time window. This suggests that there is important semantic information in both windows, in line with the results from our Time Windowing Experiment.

We see that the sources for higher accuracies in the full time window are distributed across various regions of the cortex. For further insight, we compare our results to experiments which use other imaging techniques. Work that applies these models in fMRI found the highest scoring voxels were distributed across the cortex [27, 33]. Sudre et al. found in their MEG-based results that many areas of the cortex contributed to semantic representations, but that most were parietal, occipital

or temporal [38]. When we consider later windows in time (Figure 5.5, B) the sensor distribution is more confined to parietal, occipital and temporal regions. When we consider earlier windows in time (Figure 5.5, A) the sensor distribution is more left lateralized and partially frontal. However, when we consider the full 1000ms, we see significance roughly distributed. Our results align with these studies and collectively provide evidence that the semantic representation of words is distributed across the cortex.

6.7 Future Work

The work in this thesis has led to the identification of a number of areas which are of particular interest for expanding this research. Although mentioned where appropriate in each section, here we will collect all of the avenues and elaborate on each.

While the results of the Participant Performance Experiment aligned with our hypothesis, the inability to perform a statistical significance test means we cannot definitively confirm that individual participant performance correlates directly to 2 vs. 2 accuracy. Additionally, because this experiment requires segregating participants into separate groups, the 2 vs. 2 accuracy is negatively affected. It would be valuable to explore alternative ways of identifying whether or not there is a statistically significant relationship between participant task accuracy and 2 vs. 2 accuracy with another experiment.

In the Time Windowing Experiment results we identified a statistically significant spike in 2 vs. 2 accuracy in an earlier time window that we did not expect to see. After further research, we identified a number of cases in other work where semantics have been identified in similarly earlier periods, including in the similar experiment by Sudre et al. using MEG [38]. It would be valuable to explore what leads to this early identification of statistically significant accuracy, and whether or not it contributes to similar features of the semantic word vectors when compared against the later peak in accuracy.

In a similar fashion, it would be interesting to explore in general what sections of the cortex are contributing to what features of the semantic word vectors. As the earlier research in MEG used more simple, human created word vectors they were able to identify which areas of the brain led to highly predicted values of each word vector index. The Skip-Gram word vectors are not as directly interpretable, but methods

such as Principle Component Analysis (PCA) exist which use linear transformations to reduce the dimensionality of large vectors down to a smaller set of new variables which may be more interpretable [15]. We can apply PCA to the trained model's weights, and calculate loadings from the principle components to identify features which are influencing the model similarly to other features. Since our model features represent a given sensor at a given point in time, it may provide insights to areas and timepoints of the brain's processing which contribute similar information to the overall model.

Another interesting aspect would be the comparison of words on a more individual level. Does the model tend to confuse some words more than others? While we hypothesize this is likely, it would be valuable to determine exactly which words and how semantically similar those words are.

Lastly, this research focuses on the translation of single words and on the application of this methodology to learning, but much work has been done in fMRI and MEG using sentences, adjective-noun pairs, and other more complex components of speech. With this baseline evidence that EEG is an effective means for studying semantic representation in the brain, it would be valuable to continue to replicate these experiments in EEG and further explore what benefits can be provided by this collection modality.

To work toward these goals we have initiated a few projects of interest. Our first project is a similar study utilizing EEG, but with a focus on English words. In this experiment we've chosen the original 60 words used by Mitchell et al. [27] with a larger and consistent number of repetitions per word for a larger number of participants. We anticipate this may help us continue to explore the adaption of these experiments to EEG in a more close replication of the earlier studies.

We have also initially explored the study of more cost effective EEG devices, such as the UltraCortex (Markv IV, OpenBCI, New York, USA). These style of devices are substantially cheaper and should much allow larger quantities of data collection, but consequently have the drawback that the data is poorer quality with less sensors. We hope that this approach will show some attributes of semantics remain identifiable by offsetting the more noisy signal with larger quantities of data.

6.8 Conclusion

Our experiments have identified some novel results which lead to conclusions about both semantic research using EEG and the measurement of learning in the brain. As mentioned in the last chapter, we are able to detect semantic representation using EEG and measure learning of the artificial language mapping using the same methodology.

The nature of the semantic representation experiment has improvements over prior work such as diversifying to symbols rather than words, including a larger variety of parts of speech, and the addition of more semantic categories. Our approach to measuring learning can offer improvements over the traditional reward positivity by measuring the actual concept to be learned rather than an ERP correlate. These include the ability to detect retention of learning and assess the speed of learning. We also find interesting conclusions in our time window analysis and sensor analysis experiments, including the identification of an approximately quarter second delay over prior work when introducing the translation task and the identification of a distributed nature to the semantic representation over the cortex.

This chapter completes the coverage of our research in detail. The next chapter summarizes the thesis and our contributions to the state-of-the-art.

Chapter 7

Conclusions

This chapter brings us to the end of the thesis. Through the previous chapters, we have introduced the research area of semantic representation in the human brain and covered the key previous work that established the current state-of-the-art. We've covered the experiment paradigm in this thesis: a high quality dataset of participants performing a reinforcement language learning task in which they map symbols to English words and sentences. The model that we utilize in this thesis has been covered in detail as well, and the novel experiments that we can perform by combining the previous methodolgy used in fMRI and MEG, the reinforcement learning experiment paradigm, and the valuable features of EEG.

Our contributions confirmed that we can 1) detect the semantic representation of English words in EEG while participants read the symbol language, 2) measure how these semantic representations develop over time during the participant learning phase, 3) validate and compare our model against a traditional reward positivity analysis of the same experiment, 4) provide supportive evidence that suggests intuitive alignment with the participants' task accuracies, 5) identify a delayed peak in the strength of the semantic representation correlating to the delay required for the task, and 6) provide further evidence that the source of semantic representations in the human brain is highly distributed and not simply attributable to a single area of the cortex. This research iterates on past research in many ways by providing further support for existing evidence, but the largest contributions to the state-of-the-art are that EEG is a definite valuable and affordable tool for performing semantic representation research even in complex paradigms and that this methodology is a powerful approach for analyzing learning in the human brain.

EEG can be used to detect word semantics, even for a symbol-based artificial

language, and even during the process of learning. EEG is a more noisy brain imaging modality, but even still, we provide evidence that EEG can produce decoding results on par with previous work [27]. We provided ERP, behavioral, time, and sensor based evidence that our approach models the semantics of the symbol mapping. Our work brings new understanding to the dynamics of semantic representations in the brain, and provides evidence that we can detect semantic representations *as they are learned*. This hints at several new directions for studying brain function, and the neural underpinnings of learning. In particular, our work is further evidence that EEG is an effective tool for studying the brain's mechanisms for semantics and learning, and that even fine-grained semantic distinctions can be detected using EEG.

Appendix A

Word Symbol Mapping

This appendix enumerates the complete list of words used in this thesis, and the randomly associated symbols corresponding to each word. There are 60 total words.


I


YOU


WE


BE


HAVE


GO


HAPPY


SHIRT


MARKET


SAD


BOOK


CINEMA


TIRED

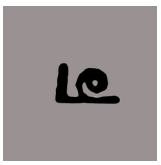

RING


HOME


HUNGRY


MONEY


FISHING


ANGRY


PENCIL


SCHOOL

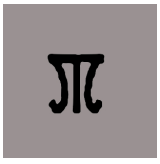

TEACHER


PHONE


LIBRARY


ARTIST


PAPER


DANCE


MUSICIAN

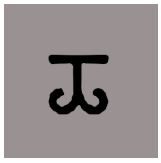

CATS


WORK


THIRSTY


SHOES


RESTAURANT


SALESMAN


EDUCATION


SHOPPING

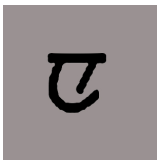

RICH


COMPUTER


SWIM


POOR


MAP


RUN


SMART


CAKE


DOWNTOWN


KIND

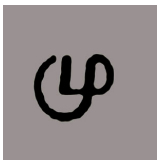

ENERGY


TRAVEL


YOUNG


CAR


CAMPING


LATE


MIRROR


HIKE


EAGAR


POWER


SKATE


IMPORTANT


PLAN


NORTH

Bibliography

[1] Horacio A Barber, Leun J Otten, Stavroula-Thaleia Kousta, and Gabriella Vigliocco. Concreteness in word processing: Erp and behavioral effects in a lexical decision task. *Brain and language*, 125(1):47–53, 2013.

[2] Christian Bellebaum and Marco Colosio. From feedback-to response-based performance monitoring in active and observational learning. *Journal of cognitive neuroscience*, 26(9):2111–2127, 2014.

[3] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

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

[5] Kai-min Kevin Chang, Vladimir L Cherkassky, Tom M Mitchell, and Marcel Adam Just. Quantitative modeling of the neural representation of adjective-noun phrases to account for fMRI activation. *In Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, Suntec, Singapore, 2-7 August 2009*, (August):638–646, 2009.

[6] David Cohen and B Neil Cuffin. Demonstration of useful differences between magnetoencephalogram and electroencephalogram. *Electroencephalography and clinical neurophysiology*, 56(1):38–51, 1983.

[7] William Croft and D Alan Cruse. *Cognitive linguistics*. Cambridge University Press, 2004.

[8] Jenna Cunningham, Trent Nicol, Steven Zecker, Nina Kraus, et al. Speech-evoked neurophysiologic responses in children with learning problems: development and behavioral correlates of perception. *Ear and Hearing*, 21(6):554–568, 2000.

[9] Anthony Christopher Davison, David Victor Hinkley, et al. *Bootstrap methods and their application*, volume 1. Cambridge university press, 1997.

[10] Alona Fyshe. *Corpora and Cognition: The Semantic Composition of Adjectives and Nouns in the Human Brain*. PhD thesis, Carnegie Mellon University, 2015.

[11] Yuqiao Gu, Fabio Celli, Josef Steinberger, Andrew James Anderson, Massimo Poeiso, Carlo Strapparava, and Brian Murphy. Using Brain Data for Sentiment Analysis. *Journal for Language Technology and Computational Linguistics*, 29(1):79–94, 2014.

[12] Felix Hill, Roi Reichart, and Anna Korhonen. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 2016.

[13] Geoff Hollis, Chris Westbury, and Lianne Lefsrud. Extrapolating human judgments from skip-gram vector representations of word meaning. *The Quarterly Journal of Experimental Psychology*, 70(8):1603–1619, 2017.

[14] Clay B Holroyd and Michael GH Coles. The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychological review*, 109(4):679, 2002.

[15] Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Phil. Trans. R. Soc. A*, 374(2065):20150202, 2016.

[16] Charles W Kreidler. *Introducing english semantics*. Routledge, 2002.

[17] Olav E Krigolson, Cameron D Hassall, and Todd C Handy. How we learn to make decisions: rapid propagation of reinforcement learning prediction errors in humans. *Journal of cognitive neuroscience*, 26(3):635–644, 2014.

[18] Olav E Krigolson, Lara J Pierce, Clay B Holroyd, and James W Tanaka. Learning to become an expert: Reinforcement learning and the acquisition of perceptual expertise. *Journal of Cognitive Neuroscience*, 21(9):1833–1840, 2009.

[19] Olave E Krigolson, Chad C Williams, Angela Norton, Cameron D Hassall, and Francisco L Colino. Choosing muse: Validation of a low-cost, portable eeg system for erp research. *Frontiers in neuroscience*, 11:109, 2017.

[20] Gina R Kuperberg. Neural mechanisms of language comprehension: Challenges to syntax. *Brain research*, 1146:23–49, 2007.

[21] Marta Kutas and Steven A Hillyard. Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427):203–205, 1980.

[22] Steven J Luck. *An introduction to the event-related potential technique*. MIT press, 2014.

[23] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[24] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[25] Wolfgang HR Miltner, Christoph H Braun, and Michael GH Coles. Event-related brain potentials following incorrect feedback in a time-estimation task: evidence for a generic neural system for error detection. *Journal of cognitive neuroscience*, 9(6):788–798, 1997.

[26] Tom Mitchell, Rebecca Hutchinson, Marcel Just, Sharlene Newman, Xuerui Wang, Radu Stefan Niculescu, and Francisco Pereira. Machine Learning of fMRI Virtual Sensors of Cognitive States. *Magnetic Resonance Imaging*, (1):1–23, 2002.

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

[28] Rachel L Moseley, Friedemann Pulvermüller, and Yury Shtyrov. Sensorimotor semantics on the spot: brain activity dissociates between conceptual categories within 150 ms. *Scientific reports*, 3:1928, 2013.

[29] Brian Murphy, Marco Baroni, and Massimo Poesio. EEG responds to conceptual stimuli and corpus semantics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 619–627. Association for Computational Linguistics, 2009.

[30] Brian Murphy, Massimo Poesio, Francesca Bovolo, Lorenzo Bruzzone, Michele Dalponte, and Heba Lakany. EEG decoding of semantic category reveals distributed representations for single concepts. *Science*, 117:131–138, 2011.

[31] Brian Murphy, Partha Talukdar, and Tom Mitchell. Selecting corpus-semantic models for neurolinguistic decoding. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, pages 114–123. Association for Computational Linguistics, 2012.

[32] E.E. Papalexakis, A. Fyshe, N.D. Sidiropoulos, P.P. Talukdar, T.M. Mitchell, and C. Faloutsos. Good-enough brain model: Challenges, algorithms and discoveries in multi-subject experiments. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 95–104, 2014.

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

[34] Cyril R Pernet, Paul Sajda, and Guillaume A Rousselet. Single-trial analyses: why bother? *Frontiers in psychology*, 2:322, 2011.

[35] Greg Hajcak Proudfit. The reward positivity: From basic research on reward to a biomarker for depression. *Psychophysiology*, 52(4):449–459, 2015.

[36] Michael D Rugg, Michael C Doyle, and Tony Wells. Word and nonword repetition within-and across-modality: An event-related potential study. *Journal of Cognitive Neuroscience*, 7(2):209–227, 1995.

[37] Svetlana V. Shinkareva, Robert A. Mason, Vincente L. Malave, Wei Wang, Tom M. Mitchell, and Marcel Adam Just. Using fMRI brain activation to identify cognitive states associated with perception of tools and dwellings. *PLoS ONE*, 3(1):1–9, 2008.

[38] Gustavo Sudre, Dean Pomerleau, Mark Palatucci, Leila Wehbe, Alona Fyshe, Riitta Salmelin, and Tom Mitchell. Tracking neural coding of perceptual and semantic features of concrete nouns. *NeuroImage*, 62:451–463, 2012.

[39] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

[40] Matthew M Walsh and John R Anderson. Learning from experience: event-related potential correlates of reward processing, neural adaptation, and behavioral choice. *Neuroscience & Biobehavioral Reviews*, 36(8):1870–1884, 2012.

[41] Xuerui Wang and Tom Mitchell. Detecting Cognitive States Using Machine Learning 2 The Star / Plus Experiment. pages 1–11, 2002.

[42] Chad C Williams, Kent G Hecker, Michael K Paget, Sylvain P Coderre, Kelly W Burak, Bruce Wright, and Olave E Krigolson. The application of reward learning in the real world: Changes in the reward positivity amplitude reflect learning in a medical education context. *International Journal of Psychophysiology*, 2017.

[43] Haoyan Xu, Brian Murphy, and Alona Fyshe. Brainbench: A brain-image test suite for distributional semantic models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2017–2021, 2016.