Decoding Semantic Representations During Production of Minimal Adjective-Noun Phrases

by

Maryam Honari Jahromi
B.Sc., Shiraz University, 2016

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

MASTER OF SCIENCE

in the Department of Computer Science

Decoding Semantic Representations
During Production of Minimal Adjective-Noun Phrases

by

Maryam Honari Jahromi
B.Sc., Shiraz University, 2016

Supervisory Committee

Dr. Alona Fyshe, Supervisor
(Department of Computer Science)

Dr. George Tzanetakis, Departmental Member
(Department of Computer Science)

ABSTRACT

Through linguistic abilities, our brain can comprehend and produce an infinite number of new sentences constructed from a finite set of words. Although recent research has uncovered the neural representation of semantics during comprehension of isolated words or adjective-noun phrases, the neural representation of the words during utterance planning is less understood. We apply existing machine learning methods to Magnetoencephalography (MEG) data recorded during a picture naming experiment, and predict the semantic properties of uttered words before they are said. We explore the representation of concepts over time, under controlled tasks, with varying compositional requirements. Our results imply that there is enough information in brain activity recorded by MEG to decode the semantic properties of the words during utterance planning. Also, we observe a gradual improvement in the semantic decoding of the first uttered word, as the participant is about to say it. Finally, we show that, compared to non-compositional tasks, planning to compose an adjective-noun phrase is associated with an enhanced and sustained representation of the noun. Our results on the neural mechanisms of basic compositional structures are a small step towards the theory of language in the brain.

Keywords: Semantic composition; language production; brain decoding;

Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

I would like to thank:

my supervisor, Dr. Alona Fyshe, for her great mentorship, patience, encouragement and making herself available throughout my graduate studies, especially during the process of writing this thesis.

Dr. Brea Chouinard, for listening to my ideas and assisting with the final stages of this research.

my lab-mates, Chris, Dhanush, ED, Cole and Isabelle, for their support, great discussions and encouragement.

my parents, for their endless and ongoing support throughout my life journey.

my partner, Payam, for his support and encouragement during tough times.

DEDICATION

To my mother, Tahere,
for her unconditional love and support

Chapter 1

Introduction

We as human beings can speak and understand an infinite number of new sentences built from a finite set of words. From an early age, we learn to combine words with correct grammar to form phrases, sentences and beyond. But what happens in our brain when we compose linguistic structures to describe our observations, when the mental representation of concepts are retrieved, and when they are combined into syntactically and semantically correct forms? To this end, many brain imaging studies have analyzed the mental representation of individual words [1, 2, 3, 4] or simple compositional phrases during language comprehension [5, 6, 7, 8]. However, these mental representations have been less explored during utterance planning; mainly due to the difficulties in designing a controlled experiment and avoiding motion contamination by the onset of speech [9, 10, 11]. In this thesis, we investigate the mental representation of concepts during utterance planning of compositional adjective-noun phrases.

Semantics is a branch of linguistics which explores the conceptual meanings at different levels of linguistic structures such as words, phrases, sentences and beyond. Computational linguistics has introduced statistical methods to model semantics, conventionally referred to as distributional semantic models (DSM). These models are obtained from large bodies of text and are able to capture different semantic properties such as gender, animacy and edibility [12].

Recent studies of semantics in the human brain use a method called *brain decoding* which predicts the properties of the stimuli from participants' brain recordings. Brain decoding is mainly based on machine learning methods and is used to characterize the mental representation of the words in human brain. Studies show that brain decoding is able to distinguish stimulus words using semantic properties related to

size, animacy or manipulability [1, 4, 5, 13].

In this thesis, we decode the semantic properties of the words before their utterance from brain recordings. We use data from an experiment conducted by Blanco-Elorrieta et al. (2018) which compares compositional versus non-compositional picture naming tasks [11]. Participants name a coloured object on the screen while their brain activity is recorded using Magnetoencephalography (MEG). They utter minimal two-word phrases consisting of a colour-describing adjective followed by an object-describing noun (e.g. red bag). We employ existing machine learning methods to decode the semantic representation of concepts from MEG during utterance planning.

Based on the results of our experiments, we find that 1) there is enough information in MEG data to decode the semantic representation of the word that a participant is about to say; 2) the neural representation of the word's semantics becomes more salient close to its articulation. By comparing conditions with varying compositional requirements, we provide evidence suggesting that 3) compositional tasks incorporate a stronger representation of the noun and 4) compositional tasks are associated with maintaining the noun representation in the brain for longer compared to non-compositional tasks.

1.1 Thesis Organization

The thesis uses existing computational semantic models to better understand the mental processes during semantic composition and language production in human brain. Chapter 2 reviews existing computational models of semantics, ways to evaluate them, and their application to study semantics in the human brain. Also in Chapter 2, we summarize the influential studies on the semantic composition of simple adjective-noun phrases in the brain. Chapter 3 describes the details of an MEG picture naming experiment conducted by Blanco-Elorrieta et al., (2018) whose data we use for this thesis [11]. We explain the stimuli, participants, data acquisition and the conditions (including compositional and non-compositional conditions) [11]. The machine learning methodologies and the evaluation of statistical significance are covered in Chapter 4. In Chapter 5, we report our results and discuss the conclusions that can be drawn. Finally, in Chapter 6, we summarize the experiment, our findings and possible future directions.

Chapter 2

Related Work

In this chapter, we review recent studies of semantic representation in text and the human brain. We explain Distributional Semantic Models (DSM) from computational linguistics and their potential use in understanding the brain through brain imaging technologies. This thesis uses DSMs to investigate how the human brain composes higher orders of semantics by combining words.

2.1 Language Semantics in Text

Models of semantic representation are based on the hypothesis that semantically similar or related words occur in similar contexts [14]. For instance, we expect that the word "hammer" be in the same context with the words "nail", "shovel" or "wood" but it is less likely to see "hammer" with the word "tomato". In Firth's seminal work, "A Synopsis of Linguistic Theory" [15], he stated that "You shall know a word by the company it keeps", which has been the fundamental idea of computational modelling of word semantics.

2.1.1 Word Representation

Models of distributional semantics approximate the meaning of a word from its co-occurrence pattern with nearby words averaged over many examples of its usage in a large body of text. DSMs obtain this goal by associating each word in vocabulary with a real-valued vector referred to as *word vector*. When represented this way, words with similar distribution patterns will have similar or related semantics. The semantic relation of two words is measured by mathematical vector similarity metrics such as

cosine similarity. Word vectors are usually modelled in two main approaches, count-based models and predictive models. Word vectors are vital in the field of Natural Language Processing (NLP) for a variety of tasks, namely sentiment analysis, text classification, language translation, etc. Furthermore, there have been studies to combine individual word representations in order to form phrases.

Count-based Word Representation

In count-based representation models, a large body of text is used to calculate the co-occurrence pattern of the words. The co-occurrence pattern are frequency counts obtained from large bodies of text such as term-document counts (number of the times term x appeared in document d) or the word-context counts (number of times a context word y appeared near the main term x, within a window of e.g. 20 words) [16, 17, 18, 19]. After some adjustments on the frequencies, (e.g. Inverse Document Frequency (IDF), Pointwise Mutual Information), a large sparse matrix is formed which is then factorized by dimensionality reduction techniques such as singular value decomposition (SVD). Consequently, a compressed representation vector is obtained, which is easier to work with. These representation vectors are conventionally evaluated on task similarity benchmarks or NLP tasks. We discuss the evaluation of word vectors in Section 2.1.3.

Predictive Word Representation

Predictive word representation is a more recent trend in the DSMs, which optimizes the representation vectors in a semi-supervised task. Generally, given a target word, a language model is trained to predict the probabilities of the context words. Once the training is done, the learned weights of said model are used as the word vector representation. Bengio et al., (2003) were the first to demonstrate the possibility of using neural probabilistic language modeling as a method of compression, instead of matrix factorization [20]. In 2013, Mikolov et al. introduced two computationally efficient models , continuous Bag-of-words (CBOW) and Skip-gram, which improved accuracy on task similarity benchmarks at a lower training cost [22]. CBOW and Skip-gram had a considerable impact on the NLP field, mainly through the supplementary toolkit named *word2vec* which facilitates training word vectors on other corpora [21]. In this thesis, we used the Skip-gram with negative sampling to study composition in the brain [21]. Mikolov et al. used negative sampling as a technique to speed up the

training of the Skip-gram model without requiring a huge amount of training samples [21].

2.1.2 Phrase Semantic Representation

New meanings in language are composed by combining several words together. Previously, DSMs have been used to model the composed meaning of phrases from their constituent word vectors. However, constructing a comprehensive model of phrase-level representation is a difficult task as there are various types of phrases from compositional (apple juice) to idiomatic (last straw). Mitchell and Lapata, (2010) compared several predefined composition operators to form adjective-noun, noun-noun and verb-object phrase representations from their constituent word vectors [23]. They concluded simple element-wise addition and multiplication of word vectors worked as well as the more complicated functions such as the tensor product or the dilation. Yet, addition and multiplication ignore word order or context information. While several other models suggest representing words, particularly adjectives, with matrices [24, 25, 26], they are either computationally costly or restricted to low-dimensional word vectors [27]. With all these corpus-based models of semantics, the question of *how the human brain composes phrase representations* remains unanswered.

In this thesis, we investigate the mental processes involving in the production of adjective-noun phrases which describe a coloured object, i.e."red bag". In our case, the adjective, which is always a colour, only changes visual properties of the noun. Therefore, instead of using phrase semantic models, we represent each word of the phrase with a separate word vector.

2.1.3 Evaluation of Semantic Representation

DSMs are generally trained based on the existing information in the context, thus a careful evaluation is needed to compare the trained representations. Most of the evaluation schemes focus on intrinsic evaluation, where they analyze the inherent relations of the words in the vector space. These methods test nearby representation vectors against human annotated datasets of synonyms or semantically related pairs of words [17, 28, 29, 30]. More recent work by Mikolov et al., (2013) demonstrated that a linear combination of word vectors encodes certain semantic analogies [22]. Baroni et al., (2014) compared several predictive word vectors against count-based word vectors on a comprehensive benchmark of word analogy, semantic relatedness,

etc. [31] They concluded that predictive models outperformed count-based models in nearly all the examined tasks [31].

While intrinsic evaluation methods build on linguistic regularities captured by DSMs, they have been criticized for a number of reasons. Faruqui et al., (2016) argued that semantic similarity is subjective, and difficult to examine out of context [32]. They also noted the inability of single word vectors to account for polysemy (words with multiple meanings). Furthermore, research has found that performance of word vectors on the word similarity datasets is not necessarily aligned with the task-specific benchmarks [33, 34].

Another line of research suggests evaluation of word vectors on their similarity to brain-imaging data. Xu et al., (2016) introduced *BrainBench*, a system for comparing various distributional semantic vectors using brain data. The brain-imaging data was recorded from language tasks in English and Italian [35, 36]. Evaluation of word vectors in BrainBench was fast and computationally efficient. In this system, Skip-gram was one of the best-performing word vector among the 6 evaluated vectors.

Abnar et al., (2018) tested word vectors to evaluate their performance for predicting the brain's activation pattern associated with 60 concrete nouns [37]. They reported predictive word vectors beat other considered word vectors. However both these studies considered brain images corresponding to a limited set of words mostly concrete nouns, which hinders them to be comprehensive benchmarks of word vectors.

2.2 Language in the Brain

2.2.1 Brain Imaging Technologies

Recent advancements in non-invasive neuroimaging technologies have let scientists capture the structure and function of the brain without having to open the skull. As a result, scientists can ask conscious subjects to perform actions during experiments while their brain behaviour is captured. Electroencephalography(EEG), Magnetoencephalography(MEG) and Functional magnetic resonance imaging (fMRI) are several brain imaging methodologies widely used in the study of language processing in the brain. Table 2.1 gives an overview of these neuroimaging methods.

High temporal resolution of the MEG signal enables scientists to study the early responses evoked by language stimuli. MEG measures the magnetic fields caused by electrical currents in the brain. Since magnetic fields are not profoundly affected by

the skull and the scalp, MEG has a relatively reliable spatial resolution.

Table 2.1: Comparison of frequently used neuroimaging methods, adapted from [38]

Neuroimaging method	Activity measured	Direct/indirect Measurement	Temporal resolution	Spatial resolution	Portability
EEG	Electrical	Direct	∼0.001 s	∼10 mm	Portable
MEG	Magnetic	Direct	∼0.001 s	∼5 mm	Non-portable
fMRI	Metabolic	Indirect	∼1 s	∼1 mm	Non-portable

2.2.2 Learning Semantics in the Brain

Numerous neuroimaging studies investigate the question of *how the brain acquires, stores, organizes and combines semantic information.* Semantics in the brain has been traditionally studied by comparing the magnitude of neural activity between experimental conditions. However, in recent years, machine learning methods has gained popularity to detect patterns of the encoded information in the neural activity. Brain decoding in particular takes brain activity as input and outputs predictions of stimulus properties.

Single Word Semantics

Classic neurolinguistic studies suggest that the semantic processing circuits of the brain are distributed across various anatomical regions [39, 40]. When a word is presented to the human brain, its meaning is encoded by semantically related sensory-motor brain areas. For example, if the word describes an action (e.g., pick), associated motor cortex elicits increased activity [41] or if the word is associated with a coloured object (eggplant-purple), colour perception areas show an activity response.

In an early example of using machine learning to decode word semantics, Mitchell et al., (2004) classified the semantic category of written stimuli viewed by the participants [42]. They were able to distinguish 12 semantic categories from fMRI recordings of the whole brain with considerable accuracy. This influential study opened the way to investigate conceptual processes in the brain using machine learning methods.

Brain recording is costly and designing experiments that collect enough data for the large vocabulary known by the language speakers would be time-consuming and not feasible. A seminal study by Mitchell et al., (2008) suggested approximating fMRI activity of related words with no available fMRI data using an intermediate semantic feature space (IFS) [43]. In this intermediate space, each word is represented

by a vector indicating how often it is used with 25 hand-picked verbs. The vector dimensions are interpretable; therefore the trained model provides insight in how much a semantic property of the noun is correlated to a specific locus of brain.

To better understand how the semantic information emerges over time (within seconds after stimuli onset), Sudre et al., (2012) have employed MEG, which has a higher temporal resolution than fMRI [4]. While recording MEG, participants answered 20 questions about perceptual and semantic features of 60 concrete nouns. A two-stage classifier was trained first to predict an intermediate feature vector and based on this vector it was able to distinguish two new nouns that it had not seen before. Based on the observed mismatch between the time course of MEG activity and semantic decoding accuracy, Sudre et al. suggested that applying machine learning models can detect a more complex pattern of semantic information in MEG data.

Adjective-noun Phrase in the Brain

The next question is *how the brain composes higher orders of meaning by combining words together*. Research concerned with the composition of simple phrases has consistently implicated the major role of the left anterior temporal lobe (LATL). Similar effects have been also reported in the ventromedial prefrontal cortex (vmPFC) [6, 8, 44] and left angular gyrus (LAG)[7]. The ordering of effects suggests the tentative mapping of early syntactic composition to LATL and later semantic composition to vmPFC.

Bemis and Pylkkänen (2011) compared MEG activity during comprehension of simple adjective-noun phrase (e.g., red boat) to non-compositional stimuli such as a list of nouns (e.g., cup boat) or a noun paired with a gibberish string (e.g., xkq boat). They found increased compositional-related activity in LATL($\sim$225 ms) and vmPFC($\sim$400 ms) [6]. Changing the same set of stimuli to auditory modality elicited increased activity in LATL, LAG but not vmPFC [7].

The LATL has also been implicated in the processing of various compositional structures such as transparent compounds (e.g. tombstone) [45] and reversed adjective-noun sequence (e.g. cup red) [8]. Along with these composition findings, LATL was also sensitive to contextual factors of phrases in a question answering paradigm [46].

Consistent with the comprehension studies, several studies on language production tasks have reported increased activity in LATL and vmPFC in preparation of simple phrase utterances [9, 10, 11]. These patterns were observed only in the conceptual

combination of adjective-noun phrases (e.g., red tree or red cups) and not the numeral combination (e.g., two cups), suggesting that the underlying effect is more sensitive to semantic modification to numeral modification [10]. Interestingly, the LATL and vmPFC show increased activity in both English and American sign language with similar timing after stimuli onset[11]. Overall, these results not only imply the LATL involvement in language comprehension but also language production in both spoken and sign language.

Computational Models of Phrase Composition in the Brain

Brain decoding methods have also been used to study the composition of adjective-noun phrases [2, 5]. These methods map the brain images to a high dimensional space of adjectives and nouns. Occasionally, a composed space is built to represent the phrase semantic based on its constituent adjective or noun representation, as discussed in Section 2.1.2.

Chang et al., (2009) explored comprehension of written adjective-noun phrases (e.g., strong dog) in fMRI data [2]. Following the study of Mitchell et al. [43], Chang et al. defined an intermediate feature vector for each noun and adjective based on five selected verbs. They also made an additive and a multiplicative composition model as suggested by Mitchell and Lapata [47]. Comparing the regression models trained to predict intermediate vectors, they reported that multiplicative compositional vectors significantly outperformed the additive and also non-compositional single word vectors. Furthermore, Fyshe et al., (2013) conducted a brain decoding study on MEG data recorded while subjects comprehend a similar set of phrases [5]. They only explored an addition and a dilation compositional model, reporting slightly better results with dilation vectors, which emphasize keeping the properties of the noun with minimal manipulations from adjectives.

In an interesting experimental design, Baron and Osherson, (2011) combined the two dimensions of gender and age to form a composed stimulus for their fMRI study [49]. Participants viewed images of boy, girl, man or woman faces while focusing on one of the eight predetermined categories (male, female, child, adult, boy, girl, man and woman). Each of the images inherently carries two dimensions: gender and age (e.g., "boy" is a combination of child and male). Baron and Osherson constructed multiplicative and additive conceptual models to approximate brain activities of the composed semantics (e.g., girl) from the activity response of their fundamen-

tal concepts (e.g., child+female). Though various regions (including LATL) showed significant additive conceptual combination, only the LATL and posterior cingulate cortex (PCC) showed multiplicative conceptual combination. Since a vast body of research identified LATL as the centre of composition, these result may reflect the multiplicative nature of compositional-related activity in LATL.

The MEG neural activities evoked by written adjective-noun phrases (e.g. tasty tomato) was explored by Fyshe et al. [5]. While the phrases were presented word by word to the participants, Fyshe et al. used brain decoding to track the semantic representation of either the adjective or the noun over time. They found that adjective representation was decodable during the presentation of the noun and even after phrase reading was completed. In this thesis, we use a similar decoding approach (discussed in detail in Chapter 4), with a different experimental paradigm.

2.3 Conclusion

Computational linguistics has proposed models of semantic representation extracted from large bodies of text. These models have helped neuroscientists decode the mental representation of semantics using machine learning methods. While brain decoding has been used to investigate comprehension of written words or minimal adjective-noun phrases, fewer studies have applied brain decoding to understand semantic processes during compositional language production. To this end, we use the data of an MEG picture naming experiment done by Blanco-Elorrieta et al. to decode semantic representation of the words while the participant is preparing to utter a compositional phrase. Before moving to the brain decoding methods, we first explain the experiment in Chapter 3.

Chapter 3

Experiment

3.1 Introduction

What happens in our brain when we describe an object by combining words? In this study, we use MEG to track the semantic composition of the words during a simple picture naming paradigm. The MEG experiment conducted by Blanco-Elorrieta et al. consists of conditions that vary the compositional requirements of the participants' utterance [11]. Their design allows us to compare semantic representations prior to utterance between compositional and non-compositional conditions. This chapter, first, discusses the experimental conditions, stimuli and the trial structures. Next, the acquisition and preprocessing of MEG data are reported. Lastly, we explain how the trials are averaged together to increase the signal to noise ratio.

3.2 Experimental Design

We use an MEG dataset collected by Blanco-Elorrieta et al. in which participants engaged in an overt naming task with pictorial stimuli (see Figure 3.1)[11]. Each stimulus was a schematic drawing of a coloured object (e.g. red bag) on a coloured background (e.g. green) and the participants are directed to describe the picture with overt speech (i.e. with minimal hesitation of doubt). The study included a target compositional condition and three non-compositional conditions. In the compositional condition, the participants were directed to name the coloured object on the screen. They uttered an adjective referring to the object colour followed by a noun describing the object's shape (e.g. red bag). This condition is compositional since

it combines a colour adjective and a noun to describe the depicted coloured object. In the other three conditions, participants were asked to name the object-name only (e.g. bag), the object colour only (e.g. red) or enumerating the list of background colour followed by the object-name (e.g. green bag). See Section 3.2.2 and Figure 3.1 for more details on the conditions and their compositional requirements.

While the task direction varies the compositional requirements of the utterance between the conditions, the stimuli picture remain constant for all the conditions. This effective design controls the perceptual processing of the stimuli with consistent pictures. Therefore, the observed difference between the conditions can only be attributed to the cognitive process involving language composition [11].

Blanco-Elorrieta et al. initially analyzed the averaged magnitude of the neural activity and concluded that during the compositional condition certain areas of the brain (LATL and vmPFC) showed a significant increase in neural activity [11]. This thesis aims to take another approach to explore the mental representations of semantic composition. We use existing machine learning methods to decode the semantic information in the neural activity before articulation. Our machine learning method can correlate the MEG recordings to the semantic properties of the stimuli. Sudre et al. have demonstrated that the peaks in the accuracy of such machine learning model are not necessarily aligned (temporally or spatially) with the increased magnitude of the neural activity in traditional event-related studies [4].

3.2.1 Stimuli

Each condition in this experiment consists of 100 trials showing a coloured object (e.g. white lamp) on a coloured background (e.g. green). From a set of five colours (black, brown, green, red, white) and five objects (bag, bell, cane, lamp, plane), Blanco-Elorrieta et al. constructed all the 25 possible combinations that form a coloured object. For each of the coloured objects, another colour is selected for the background from the initial set of colours. The background colours are repeated the same number of times amongst the 25 pictures and equally distributed across the five object shapes. Afterwards, for each picture, Blanco-Elorrieta et al. added a new picture where the object's colour and the background were interchanged (e.g. for a *white lamp on green background*, they include *a green lamp on a white background*). Every picture is repeated 2 times forming a set of 100 pictures per condition. The same set of stimuli is randomized for each condition and participant while the task instruction

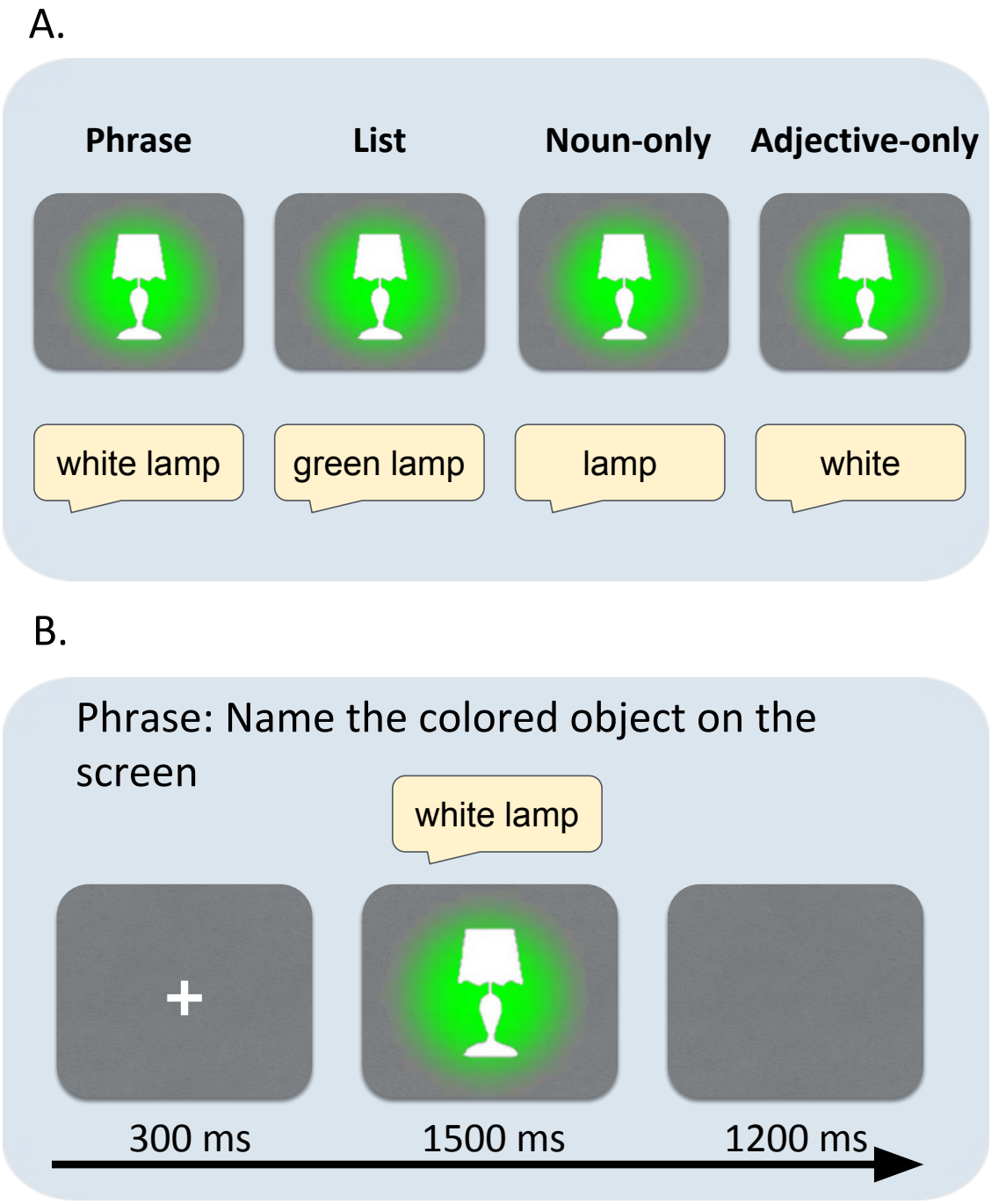

Figure 3.1: **A.** Examples of the stimulus picture and the corresponding expected utterance in the four experimental conditions. *Phrase*: object colour followed by the object-name, *List:* the background colour followed by the object-name, *Noun-only*: the object-name, *Adjective-only*: the object colour **B.** Trial structure in order: A fixation cross appears for 300 ms, then the stimulus picture presented until utterance or time out of 1500 ms. A break of 1200 follows before the start of the next trial.

was presented at the beginning of each condition.

3.2.2 Conditions

This experiment consists of four conditions, a *phrase* condition (the target condition; involves composition), a *list* condition (the non-compositional control condition), an *adjective-only* condition and a *noun-only* condition. Observing the stimuli in Figure 3.1, a white lamp on green background, participants were directed to:

- ***Phrase condition***: name the object colour followed by the object name (e.g. white lamp) which forms an adjective-noun phrase.
- ***List condition***: name the background colour followed by the object name (e.g. green lamp). Although this condition is lexically similar to the phrase condition, the adjective does not describe the properties of the noun. Thus instead of syntactically forming a compositional phrase, participants enumerate a list of an adjective followed by a noun (i.e. non-compositional list).
- ***Noun-only condition***: name the noun describing the object's shape (e.g., lamp).
- ***Adjective-only condition***: name the adjective describing the object colour(e.g. white).

The experiment is effectively designed to control the perceptual processing of the stimuli while varying the compositional requirements of the tasks across the phrase and list conditions. To achieve this, both the stimuli set and the combination of uttered parts of speech remained constant between the phrase and list conditions. Thus, increasing the likelihood that the observed differences in brain activity are attributable to the cognitive processes involving language production.

We would like to emphasize that the list condition is an unnatural language production which might feel intuitively harder due to the object-colour inhibition. Observing a coloured object, participants are forced to ignore the colour of the object, and retrieve the background colour as the adjective before naming the object shape. According to Blanco-Elorrieta et al., no reliable reaction time difference was observed between the list and phrase conditions, suggesting that inhibition was unlikely an influence in this conditional comparison.

3.3 Data Acquisition

Twenty right-handed monolingual native English speakers (9 female; ages: Mean:25.6, SD=7.3), all neurologically intact with normal or corrected to normal vision, provided their written consent to participate in this experiment conducted by Blanco-Elorrieta et al., (2018) [11]. Prior to the MEG recording, each participant's head shape was digitized by a Polhemus dual source hand-held FastSCAN laser scanner (Polhemus, VT, USA). The MEG data were recorded using a 208 channel axial gradiometer system (Kanazawa Institute of Technology, Kanazawa, Japan) at Neuroscience of Language Lab in NYU Abu Dhabi. An MEG compatible microphone (Shure PG 81, Shure Europe GmbH) was used to record uttered speech of the participants.

Each trial started with a fixation cross for 300 ms, followed by the stimuli image which was present until participant's response or timeout(1500 ms). Afterwards, a break of 1200 ms was given until appearance of the fixation cross belonging to the next trial. Although erroneous articulations including wrong naming or utterance repairs were reported, response accuracy was above 97% . The latency of speech onset for each condition was as follows: 1) List condition: Mean=897 ms 2) Phrase condition: Mean=917 ms 3) Adjective-only condition: Mean=772 ms 4) Noun-only condition Mean=792 ms.

3.4 MEG Data Preprocessing

The MEG signals were band-passed using a Butterworth filter of order 20 between 0.1Hz and 40HZ cutoff with a sampling frequency of 1000Hz. Trials were epoched at 100 ms before to 700 ms after stimuli onset to avoid contaminating motion artifact that would result from overt speech. Subsequently, epochs were baseline corrected with an average of a 100ms interval prior to the stimuli onset.

3.4.1 Averaging Epochs

Since our methodology tracks semantic representation of individual words (noun or adjective), we randomly average the epochs to cancel out noise according to a target word. Depending on the condition and the uttered parts of speech, we first choose a target and then perform the averaging for each subject separately. Assume the target word is the noun which describes the object shape. On two-word conditions (phrase

and list), we can choose the target word to be either noun or adjective. We have 100 epochs distributed equally among 5 distinct noun, each with 20 epochs. The 20 epochs of each noun are randomly divided into 4 groups, and the epochs of each group are averaged together yielding 4 *averaged epochs* per noun and 20 averaged epochs in total. We would follow the same procedure of averaging epochs if the target word were the adjectives.

3.5 Conclusion

Given the motivation to characterize the semantic representation of words during composition, we described an MEG picture naming experiment designed and collected by Blanco-Elorrieta et al.. The next chapter covers the decoding methodologies, which finds a mapping from MEG data to semantic properties of the words uttered by the participants.

Chapter 4

Methodology

The previous chapter explained the experiment design and data collection. In this chapter, we explain the existing methodology that maps the MEG data to semantic representations of the words. Section 4.1, describes the Skip-gram word vectors and their usage in decoding semantics from the brain. Next, we discuss the details of a linear regression model that predicts which of the two words is uttered by the subject from their MEG recorded data. A similar approach was followed by Fyshe et al. [5] to decode the comprehension of adjective-noun phrases. Lastly, we explain temporal generalization matrices (TGM) which test the stability of neural code over the course of time or across experimental conditions.

4.1 Prediction Framework

We use a machine learning regression model to answer the question of *when the brain encodes the background colour, object colour or object shape in preparation of a language production task*. Specifically, we use a ridge regression model, which takes the brain recordings as input and predicts the semantic properties of the stimulus word. The ridge regression model is trained and tested independently for each subject on several short windows of time prior to the utterance. For each time window, we obtain a decoding accuracy which indicates how well the semantic representations of the target words were decoded. The variation in the model accuracy over time implies how the semantic representation of the word is processed. Similarly, comparing the accuracy across the conditions implies how semantic representation is different between the phrase and the list conditions.

4.1.1 Word vectors

As summarized in Chapter 2, word vectors are real-valued vectors of numbers representing the semantic properties of a word. These vectors are approximated by predictive distributional semantic models (DSM). In this study, we make use of Skip-gram vectors [22] to identify each word in the participant's utterance. Skip-gram vectors are obtained by training an artificial neural network model. Given a word, this neural network model is trained to predict the co-occurrence probability of every other word in the vocabulary nearby the input word. The co-occurrence probabilities are calculated from the Google dataset of news articles with a vocabulary size of 692K different words. Once the training is done, the model's weights form a 300-dimensional vector which represents the semantics of the input word. Note that each dimension is not interpretable on its own. Skip-gram vectors, trained for English and other languages (Finnish and Italian), have become popular in recent studies to investigate semantic representations encoded in brain activities [50, 51, 52].

4.1.2 Prediction Model

We train multiple ridge regression models to approximate a mapping function from the recorded MEG dataset $X \in \mathbb{R}^{N \times p}$ to a multidimensional semantic space $Y \in \mathbb{R}^{N \times d}$. This semantic space Y is associated with either the adjective or the noun in the participants articulation and represents each word with a d-dimensional vector.

In Section 3.4.1, we discussed how we epoch the trials and then randomly average those with same word articulation together to form an *averaged epoch*. Each of the N averaged epochs is a $c \times t$ matrix of brain activity response collected by c gradiometer MEG channels over a time slice of t recording samples. This $c \times t$ matrix is then reshaped to a p-dimensional vector ($p = ct$) to construct a row in the input matrix $X \in \mathbb{R}^{N \times p}$. In our case, X consists of 20 averaged epochs with 208 MEG sensors and 800 recording samples over time. We normalize each column of X to have mean of 0 and standard deviation of 1. We also append a column of ones to account for the bias term of the regression model.

We learn d independent hypothesis functions $h_j(X), j \in \{1, 2, ..., d\}$ to predict the jth column of semantic matrix Y denoted as y_j^t. All the functions are trained in the same way, hence for the sake of simplicity, we represent the formulas for one hypothesis function $h_j(X)$ which maps the input matrix $X \in \mathbb{R}^{N \times p}$ to the target

column $y_j^T \in \mathbb{R}^{N\times 1}$:

$$h_j(X) = x_i^t w \tag{4.1}$$

we estimate the vector $\hat{w}$ with a linear least-squares objective criterion and l2-norm regularization (called ridge regression):

$$\hat{w} = \underset{w}{\operatorname{argmin}} \left\| Xw - y_j^T \right\|_2^2 + \lambda w^T w \tag{4.2}$$

which after derivation gives the following solution for $\hat{w}$:

$$\hat{w} = (X^T X + \lambda I)^{-1} X^T y_i^T \tag{4.3}$$

We determine the best performing regularization parameter λ separately for each column of semantic space by a leave-one-out cross validation; the most common value is $\lambda = 0.1$. In most brain decoding tasks, a small number of instances (averaged epochs) are available compared to the number of features ($N \ll p$). Therefore, inversion of a $p \times p$ matrix in equation 4.3 is computationally expensive (takes $O(p^3)$ operations). Using a kernel trick demonstrated by Hastie and Tibshirani based on singular-value decomposition (SVD), we speed up this inversion (down to $O(pn^2)$ operations) [53].

4.1.3 2 versus 2 test

To evaluate the performance of the described prediction model, we use a variation of leave-two-out-cross-validation which is called **2 versus 2 test**. On a dataset of N averaged epochs, we hold out 2 averaged epochs with target vectors (y_i, y_j) and train the model on the remained $N-2$ averaged epochs. Testing the model on the 2 held-out averaged epochs provides two predicted semantic vectors $(\hat{y}_i, \hat{y}_j)$. The 2 vs. 2 test measures how similar the predictions $(\hat{y}_i, \hat{y}_j)$ are to their corresponding ground truth vectors (y_i, y_j) using a vector distance criterion $d(v, u)$. In particular, the test passes if the following equation holds:

$$d(\hat{y}_i, y_i) + d(\hat{y}_j, y_j) < d(\hat{y}_i, y_j) + d(\hat{y}_j, y_i) \tag{4.4}$$

where the distance of matching vectors is smaller than the distance of non-matching ones. We consider a score of 1 if the test passes, 0 if it fails and 0.5 if the two summations are equal. While any kind of distance metrics can be used, we opt for

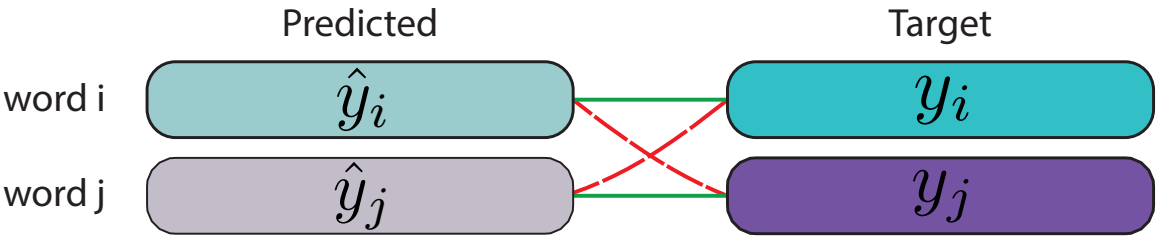

Figure 4.1: An illustration of the 2 vs. 2 test for 2 word vectors given their predicted $(\hat{y}_i, \hat{y}_j)$ and ground truth values (y_i, y_j). The 2 vs. 2 test passes if sum of matching vectors distances (solid green lines) is less than non-matching ones(dashed red lines). [See equation 4.4]

cosine distance. Finally, we report the *2 vs. 2 accuracy* which is the averaged score of the 2 vs. 2 test on every possible pair in the dataset (total of $\binom{N}{2}$).

4.1.4 Statistical Significance

To assess the statistical significance of the 2 vs. 2 accuracy, we use permutation tests. For a permutation test, the prediction framework is trained on data in which the labels are randomly shuffled after averaging epochs. In our case, the label of each epoch is the Skip-gram vector corresponding to the target stimulus word. We randomly shuffle the labels for 100 times to break any relation between the MEG recordings and semantic labels. We then fit a normal kernel density function to the 2 vs. 2 results of the 100 repetitions and form a null distribution. As we expected the mean of the empirical null distribution is at the chance level of 50%. Subsequently, the p-values of the reported results with the correct label assignment are calculated from this null distribution. We correct the p-values for multiple comparisons over time using False Discovery Rate (FDR) with no dependency assumption (Benjamini-Hochberg-Yekutieli method) [54].

We train our decoding framework over time on all 4 conditions of this study including adjective-only, noun-only, list and phrase [see Chapter 3]. We then find the clusters of time in which the 2 vs. 2 accuracy differ significantly between the two selected conditions; and only report time clusters which contain 3 or more consecutive 2 vs. 2 accuracy points. To do so, we use a separate significant test called non-parametric permutation test Maris and Oostenveld [55]. We randomly assign the condition labels to decoding results of the 20 participants for 2^{20} times and report clusters with p-value < 0.05.

4.2 Temporal Generalization Matrix (TGM)

The Temporal Generalization Matrix (TGM) is a method to assess whether a mental pattern is stable or changing over time or across conditions [56]. We test the same prediction framework described in Section 4.1 for its ability to generalize the learned weight maps to another time window. Precisely, instead of only training and testing on the same time window, we form a matrix M where M_{ij} contains accuracy of the prediction model trained on a window centred at time i and tested on another window centred at time j. To perform the 2 vs. 2 tests, we first leave two averaged epochs out from both time windows and train the weight maps on time window i using the $N - 2$ remaining words. We then test the learned weight maps on the two left-out averaged epochs from time window of j. If the mental representation of the word is consistent over time, then the trained and tested models on different time windows will show similar accuracy.

For instance, Figure 4.2 illustrates a TGM for the phrase condition averaged over all subjects. Each cell shows the results for training and testing our prediction framework on windows of 100 ms with a sliding step of 50 ms. While the diagonal cells refer to training and testing on the same time window, off-diagonal cells reveal the resemblance of the mental representations between different time windows. Another possibility is to change the experimental conditions between train and test datasets. In our example, for every time window, we could substitute the training dataset with list condition and study the generalization of the list condition on every time window in the phrase condition.

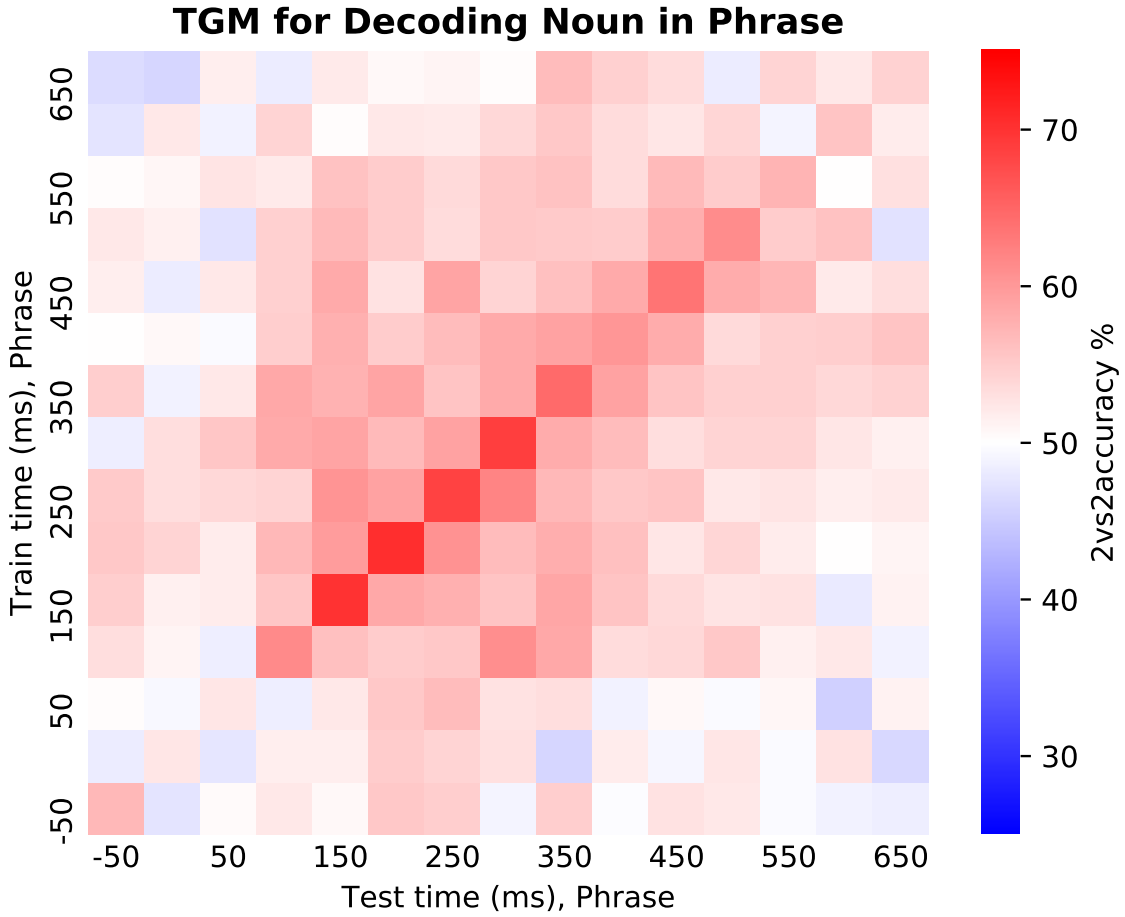

Figure 4.2: A Temporal Generalization Matrix (TGM) for decoding noun by training and testing over different time windows in the phrase condition. The diagonal line of the matrix refer to training and testing on the same window of time.(TGM results discussed in Section 5.4)

4.3 Conclusion

This chapter described how a computational model could be trained to decode the semantic representation of words using MEG data collected during a picture naming task. The computational model evaluates the amount of correlation between MEG data and semantic word vectors. In chapter 5, we report the results of the decoding models and discuss the inferences made about the mental preparation of compositional phrases before speech production.

Chapter 5

Results and Discussion

Our methodology, discussed in Chapter 4, measures the correlation of MEG signals to the semantic properties of the articulated words. In this chapter, we explore the time course of this correlation taking advantage of MEG's temporal resolution. Section 5.1 covers single word conditions, and shows that it is possible to decode the semantics of the adjective or the noun used in isolation before the utterance. Section 5.2 compares adjective and noun decodability in the phrase and list conditions to their decodability in the single word conditions. In later sections, we investigate how well the trained model on a certain condition and time window generalizes to another condition and time window. We found higher decoding accuracy of the noun in the phrase condition, which suggests that composition enhances the neural representation of the noun. We also observe a rise in decoding accuracy of the word closer to the utterance, which is associated with speech planning of the word that the participant is about to say.

5.1 Decoding Semantics in Single Word Conditions

Previous studies have predicted the semantic representation of words from brain imagining data recorded during the comprehension of written stimuli [4, 3, 5]. Building on the same methodology, we show that semantic decoding is also possible before a person says the word. As mentioned in Section 4.1.1 , we use Skip-gram vectors as the semantic representation of the words uttered by participants. Initially, we applied the decoding model on the adjective-only and noun-only conditions to analyze the mental representations of the words in isolation. Stimulus is presented at 0 ms and MEG data is truncated at 700 ms to avoid speech contamination. Training our machine

learning model using recordings of all 208 MEG electrodes from stimulus onset to 700 ms gives 2 vs. 2 accuracy of **70.53%** on decoding the adjective in the adjective-only condition and **73.62%** on decoding the noun in the noun-only condition. Chance accuracy is 50%, and permutation tests indicate that both adjective and noun decoding accuracy are significantly above chance level with $p\text{-}value < 0.05$; For details of the permutation test, refer to Section 4.1.4.

The high temporal resolution of MEG data also allows us to investigate the changes in decoding accuracy over time during the speech planning. We train our decoding model on windows of 100 ms with a 5ms sliding step. We use this setup for every temporal figure in this chapter. The variations in the accuracy of decoding the noun in the noun-only condition and the adjective in the adjective-only condition are illustrated in Figure 5.1. The significant accuracies are determined by permutation tests and FDR corrected over time (see Section 4.1.4).

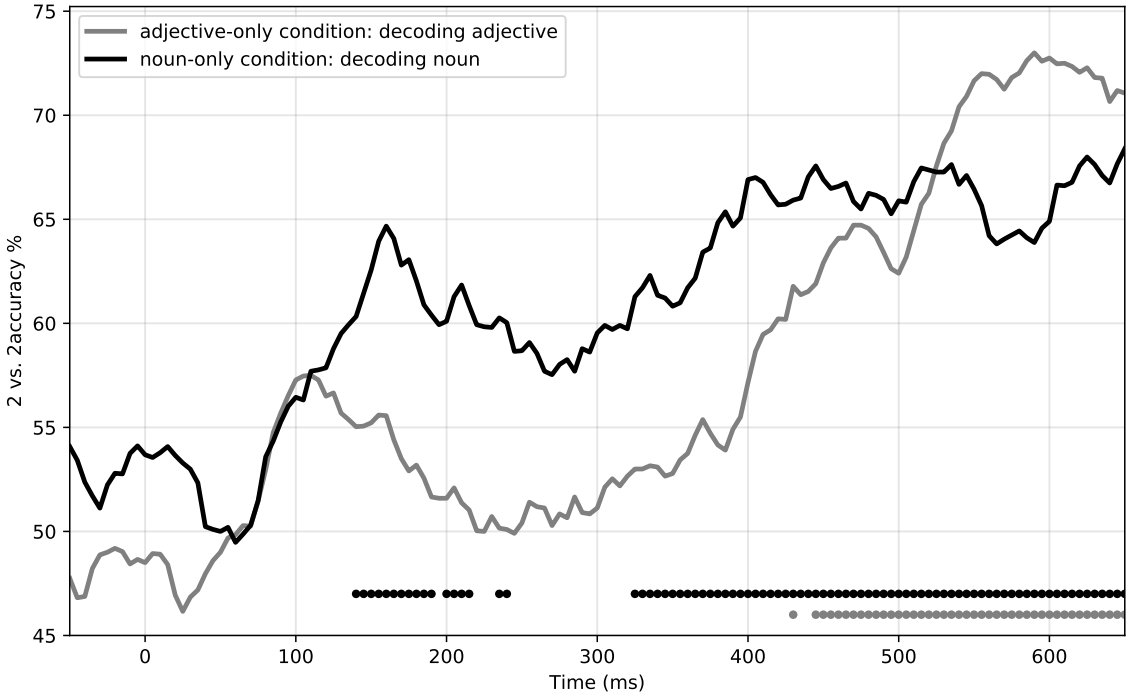

Figure 5.1: The 2 vs. 2 accuracy of decoding noun in the noun-only condition (black) and adjective in the adjective-only condition (grey). Each point represents a model that is trained on a 100 ms window (50 ms before, 50 ms after), starting from -100 ms with 5 ms sliding step. Stimulus onset is at 0 ms. Chance accuracy is 50 % and statistically significant points are denoted with dots at the bottom of the graph ($p < 0.05$, permutation tests, FDR corrected for multiple comparisons over time).

Based on Figure 5.1, the adjective shows a gradually rising accuracy which becomes significant about 430 ms after seeing the stimulus and remains above the FDR threshold until the end of the time course with the highest accuracy of **73.29%** at ∼590 ms. The 2 vs. 2 accuracy of the noun begins to be significant at around 130 ms for a time span of ∼50 ms. After dipping below the FDR threshold briefly for a couple of times, it rises again at about 325 ms and remains consistently significant until the last analyzed time window centred at 650 ms. Recall that in these conditions (adjective-only and noun-only), participants say only one word, and decoding results in both conditions are reliably above chance for at least 200 ms toward the end of recording time right before the utterance.

It is important to note that Figure 5.1 only indicates the time windows in which the mental representation of the words is significantly decodable. Based on these results, we still cannot draw any conclusion about the stability of mental representations across time, but we will cover this later in Section 5.4.1.

5.2 Decoding Semantics in List & Phrase Conditions

In this Section, we analyze the semantic representation of adjectives and nouns in the phrase and the list conditions. Both of these conditions involve the utterance of an adjective followed by a noun. However, they differ in the compositional requirements of the task. The phrase condition is compositional because the uttered colour adjective is describing the noun; whereas, in the non-compositional list condition, the adjective describes the background colour which has no relation to the noun. The stimuli set and the combination of uttered words are the same for both conditions; hence the observed difference between the conditions can only be attributed to the varying compositional requirements.

We first train a machine learning model to decode the semantics of the adjective and the noun from MEG data of 0 to 700 ms, with results reported in Table 5.1. Chance accuracy is 50%, and permutation tests indicate that all decoding accuracies are significantly above chance ($p\text{-}value < 0.05$). For decoding the adjective, the list condition has a higher 2 vs. 2 accuracy compared to the phrase condition. Whereas, for decoding the noun, the phrase condition shows a higher accuracy compared to the list condition.

Table 5.1: The 2 vs. 2 accuracy of decoding semantics of the adjective and the noun on MEG samples of 0-700 ms. The chance accuracy is 50% and all the numbers are statistically significant with p-values< 0.05

	Adjective	Noun
List	71.03	62.97
Phrase	64.41	70.5

We can see the decoding accuracy of the adjective and the noun over time in Figure 5.2.A and 5.2.B respectively. In the list condition, adjective decodability exceeds the chance level early at 65 ms peaking at about 100-200 ms and staying above chance until 245 ms. The 2 vs. 2 accuracy of the adjective increases again towards the end of time window around 550 ms. In contrast to the list condition, adjective results for the phrase condition are not significant until $\sim$595 ms.

As indicated by the highlighted areas of Figure 5.2.A, the 2 vs. 2 accuracy of the adjective is significantly higher in the list condition than the phrase condition during the early time window of 70-185 ms. This significant difference might reflect a difference in ease of colour perception between the list and phrase conditions. Recall that in the list condition, the adjective is referring to the background colour; however, in the phrase condition, it is referring to the object colour. Since the background colour covers a larger area that is mainly consistent across trials, it may be perceptually easier to detect. If so, then the perceptual dominance of the background colour might be a reason for the early peak in decoding accuracy of the adjective in the list condition. Another hypothesis for this difference involves the inhibitory nature of the list condition, where looking at a coloured object (e.g. white lamp), participants are required to inhibit the colour of the object in order to name the background colour before naming the noun (e.g. green lamp in Figure 3.1). This unnatural task might encourage participants to focus on encoding the background colour early in time resulting in a clear adjective representation and higher decoding accuracy in the list condition compared to phrase condition.

Adjective decoding for both list and phrase conditions appear to rise toward the end of the chosen time span. The MEG data is truncated at 700 ms, which is before any speech movement contaminates the data. If we could continue recording during the utterance, perhaps the decoding accuracy of the adjective would remain significant. We hypothesize that the neural code incorporates semantic properties of the word close to its articulation.

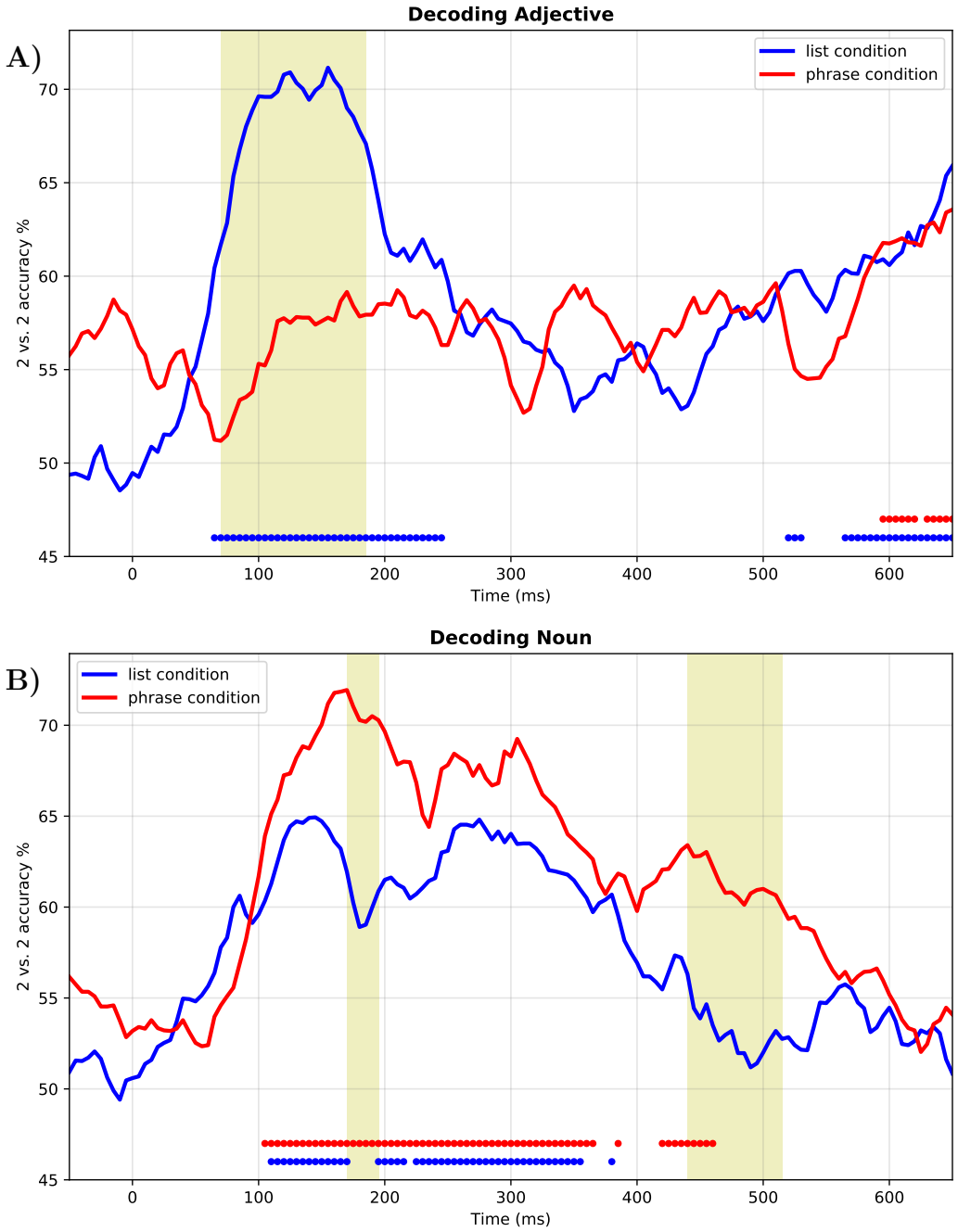

Figure 5.2: The 2 vs. 2 accuracy of semantic for decoding the A) adjective and B) noun over time in list (blue) and phrase (red) conditions. Each point represents a model that is trained on a 100 ms window (50 ms before, 50 ms after), starting from -100 ms with a 5 ms sliding step. Stimulus onset is at 0 ms. Chance accuracy is 50% and Statistically significant points are denoted with dots at the bottom of the graph ($p < .05$, permutation tests, FDR corrected for multiple comparisons over time). Shaded areas denote significantly different clusters of time between the conditions ($p < .05$, non-parametric cluster-based permutation [55]).

Concerning the noun decodability (see Figure 5.2.B), list and phrase conditions peak about the same time around 150 ms after stimuli onset. However, the 2 vs. 2 accuracy of the noun in the phrase condition is more stable (105-385 ms) and resurges above chance at ∼420 ms till ∼485 ms. This implies that the phrase compositions requires maintaining the semantics of the noun in mind for an extended time.

As the onset of the speech nears, unlike adjective, the noun decoding accuracy gradually degrades for both conditions. In both phrase and list conditions, the colour adjective is uttered first, which may explain the rising pattern of accuracy we see in later time windows in Figure 5.2.A, immediately prior to utterance. The rising accuracy of the adjective and declining accuracy of the noun supports our hypothesis of enhanced semantic representation close to utterance. Replicating this study with reversed compositional structures (e.g. noun-adjective phrases in other languages) could further delineate this phenomenon.

5.3 Generalization across Conditions

We were also interested in comparing how much the neural representation of words differed across the conditions. To achieve this, we train our decoding model on one condition and perform the 2 vs. 2 test using data from another condition. Figure 5.3 reveals that the 2 vs. 2 accuracy for training on the noun-only condition and testing on phrase and list conditions. The prediction model learned on the noun-only condition appears to generalize well to both phrase and list conditions. This generalized accuracy starts to rise early after stimulus onset (∼100ms), peaking at 150 ms. There are three windows of time in which phrase accuracy is significantly higher than list accuracy (see shaded areas). In the first and earliest window (145-210ms), the 2 vs. 2 accuracy is significant for both list and phrase condition, in the second window (320-410 ms), only the phrase accuracy is significant and in the final window (425-475 ms), neither of the conditions are above FDR threshold at any point.

Altogether, Figure 5.3 indicates that a decoding model learned from the noun-only condition generalizes better to the phrase than the list condition. The results show that the peak accuracy of this model (73.22%) is even numerically higher than models trained and tested within phrase (71.94%) or noun-only conditions (64.67%). This implies that composition in the adjective-noun phrase enhances the representation of the noun.

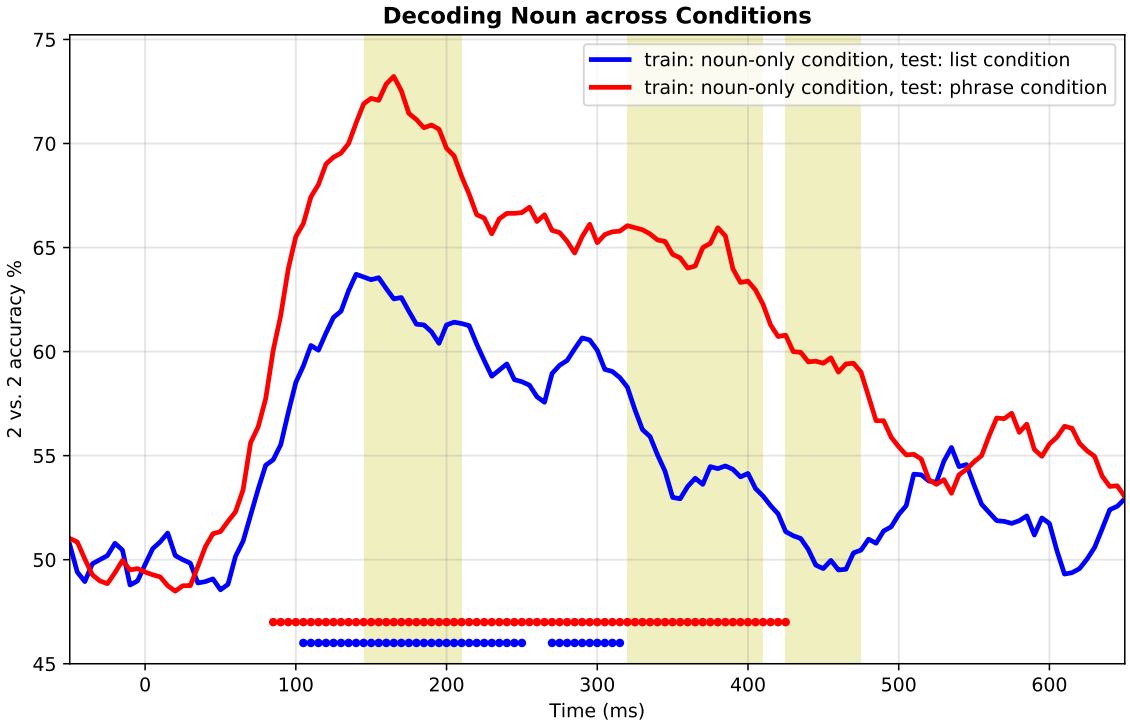

Figure 5.3: The 2 vs. 2 accuracy for decoding the noun semantics. Training on noun-only condition and testing on list (blue) and phrase (red) conditions. Each point represents a model that is trained and tested on the same 100 ms window of time between the conditions (50 ms before, 50 ms after) with a 5 ms sliding step. Chance accuracy is 50% and Statistically significant points are denoted with dots at the bottom of the graph ($p < .05$, permutation tests, FDR corrected for multiple comparisons over time). Shaded areas denote significantly different time clusters between conditions ($p < .05$, non-parametric cluster-based permutation [55]).

We exploit the same cross-condition framework to decode the adjective across the conditions. Figure 5.4 illustrates the results of training on the adjective-only condition and testing on the phrase and list conditions. We first train our machine learning model on the adjective-only condition to predict semantic properties of the adjective. Testing this model on the phrase condition shows an early statistical significant peak centered at ∼150 ms and a later gradual surge which exceeds the significance threshold from about 500 ms onward. However, testing the model learned from the adjective-only condition on the list condition does not significantly predict adjective at any time point. As illustrated by the shaded area, the early peak of testing on the phrase condition is significantly higher than testing on the list condition from 140 ms to 175 ms.

It is important to note that the referent of the adjective in the list condition differs from the phrase and the adjective-only conditions. That is, in the list condition, the adjective refers to the background colour, while it refers to the object colour in the other two conditions. This might explain why the adjective representation learned from the adjective-only condition generalizes to the phrase condition but not to the list condition.

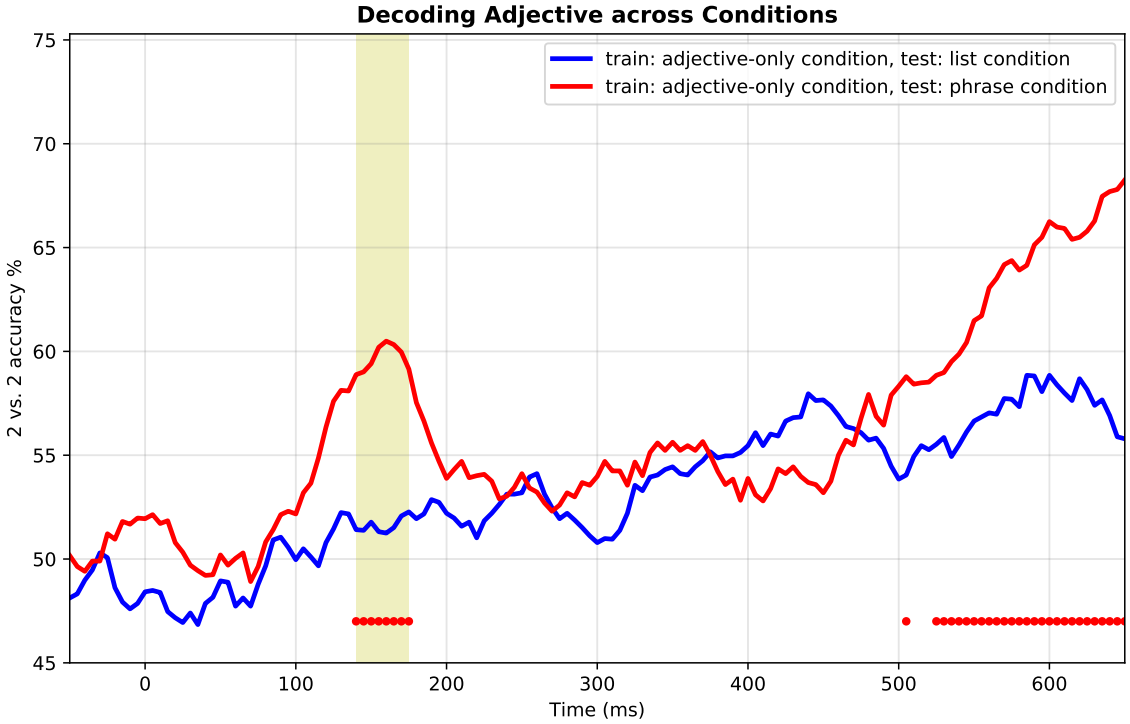

Figure 5.4: The 2 vs. 2 accuracy for decoding the adjective semantics. Training on adjective-only condition and testing on list (blue) and phrase (red) conditions. Each point represents a model that is trained and tested on the same 100 ms window of time between the conditions (50 ms before, 50 ms after) with a 5 ms sliding step. Chance accuracy is 50% and Statistically significant points are denoted with dots at the bottom of the graph ($p < .05$, permutation tests, FDR corrected for multiple comparisons over time). Shaded areas denote significantly different time clusters between conditions ($p < .05$, non-parametric cluster-based dpermutation [55]).

5.4 Temporal Generalization Matrix (TGM)

So far, we trained and tested our models on the same windows of time to decode the semantic representation of the adjective or the noun. We then extended the

decoding paradigm across the conditions. However, it is not clear how consistent the word's mental representation is over time. For example, in the list condition, is the early representation of the adjective similar to its late representation close to the utterance? Or at which time points the noun representation is consistent between two different conditions? As explained in Section 4.2, we use temporal generalization matrices (TGMs), outlined by King and Dehaene (2014), to assess the stability of words' representation over time and their similarity between the conditions.

Previously in Figure 5.2.A for decoding the adjective in the list condition, we observed two time intervals with above chance 2 vs. 2 accuracy: an early peak (from ∼65 ms to ∼245 ms) and a later gradual rise (∼550 ms to 650 ms). We were interested to understand whether the mental representation of adjective during the first above chance time interval is similar to its representation in the later time interval. To explore this stability, we tested the generalization of the adjective in the list condition using the TGM depicted in Figure 5.5. In this TGM, there are no significant off-diagonal cells located at either (train:∼150, test:∼600) ms or (train:∼600, test:∼150) ms. Based on this TGM then, we cannot conclude that the mental representations of adjective in the early and the late time intervals are similar. More investigation is needed to uncover the nature of mental processes in these two time intervals.

5.4.1 TGMs for Inter-condition Decoding of the Noun

In this section, we investigate the cross-condition generalization of the noun. Figure 5.6.A illustrates the TGMs for generalization of noun semantics between the noun-only and phrase conditions, and Figure 5.6.B shows the noun generalization between the noun-only and the list conditions. The train datasets are represented along the vertical axis, and the test datasets are along the horizontal axis. To refer to each TGM easily, we name the train condition followed by the test condition, i.e. the noun-only/phrase TGM (Figure 5.6.A, bottom-right) is trained on the noun-only condition and tested on the phrase condition. Note that the diagonal line in each TGM corresponds to training and testing on the same time windows.

Both the noun-only/noun-only (Figure 5.6.A, bottom-left) and the phrase/phrase TGMs (Figure 5.6.A, top-right) reveal significantly above chance accuracy patches around the diagonal. In order to understand whether the observed significant cells in these two TGMs are decoding similar mental representations, we interchanged

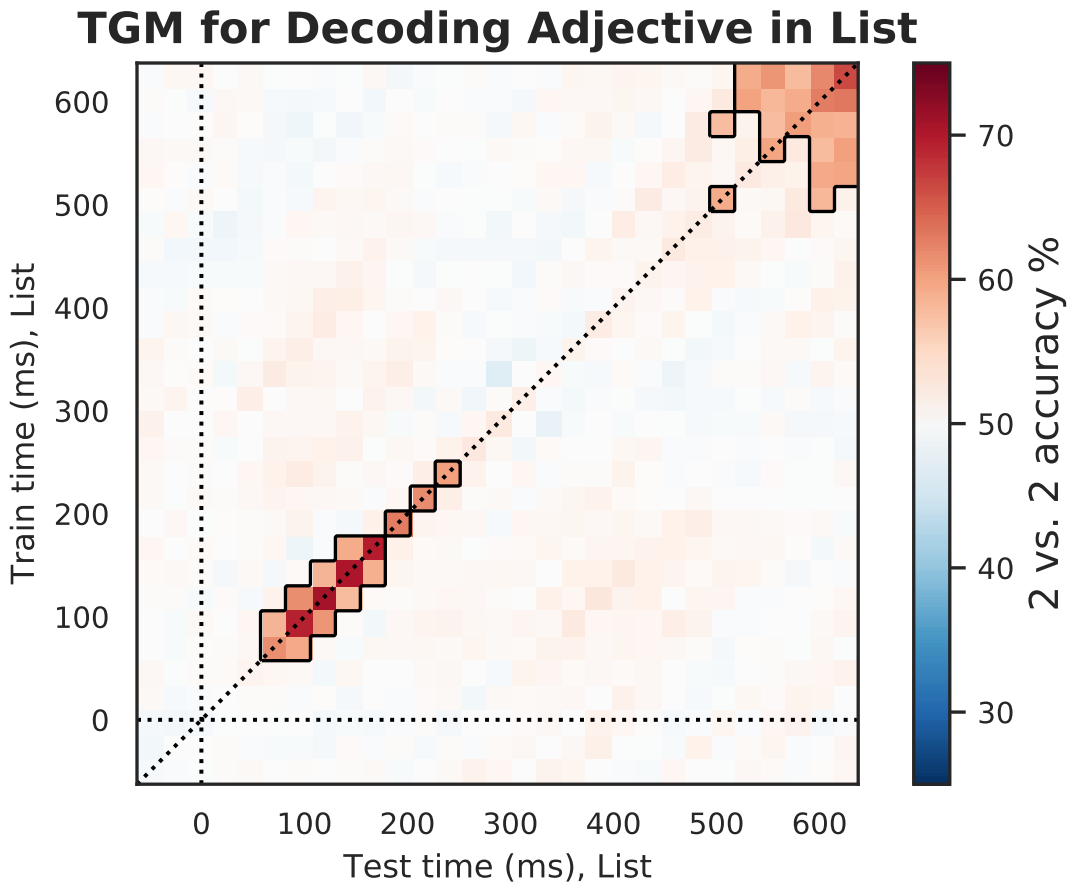

Figure 5.5: Temporal Generalization Matrix (TGM) for decoding adjective in the list condition with stimulus onset at 0 ms. The training time windows are placed along the vertical axis, and the testing time windows across the horizontal axis. All time windows are 100 ms long, starting from -100 ms with a 25 ms sliding step. Each decoding model is trained on a selected time window and tested on every available time window of 100 ms. Cells on the diagonal dashed line corresponds to training and testing on the same time window. Cells with significant accuracy are contoured ($p < 0.01$, FDR corrected for multiple comparisons over training and testing time). The early and the late diagonal significant patches (∼150 ms and ∼600 ms) do not generalize to each other.

the training and testing data. We can see the generalization results in the noun-only/phrase (Figure 5.6.A, bottom-right) and the phrase/noun-only TGMs (Figure 5.6.A, top-left) which both show a significant diagonal generalization between the conditions. The noun-only/phrase TGM (Figure 5.6.A, bottom-right) has a significant patch below the diagonal which indicates that models trained on the noun-only

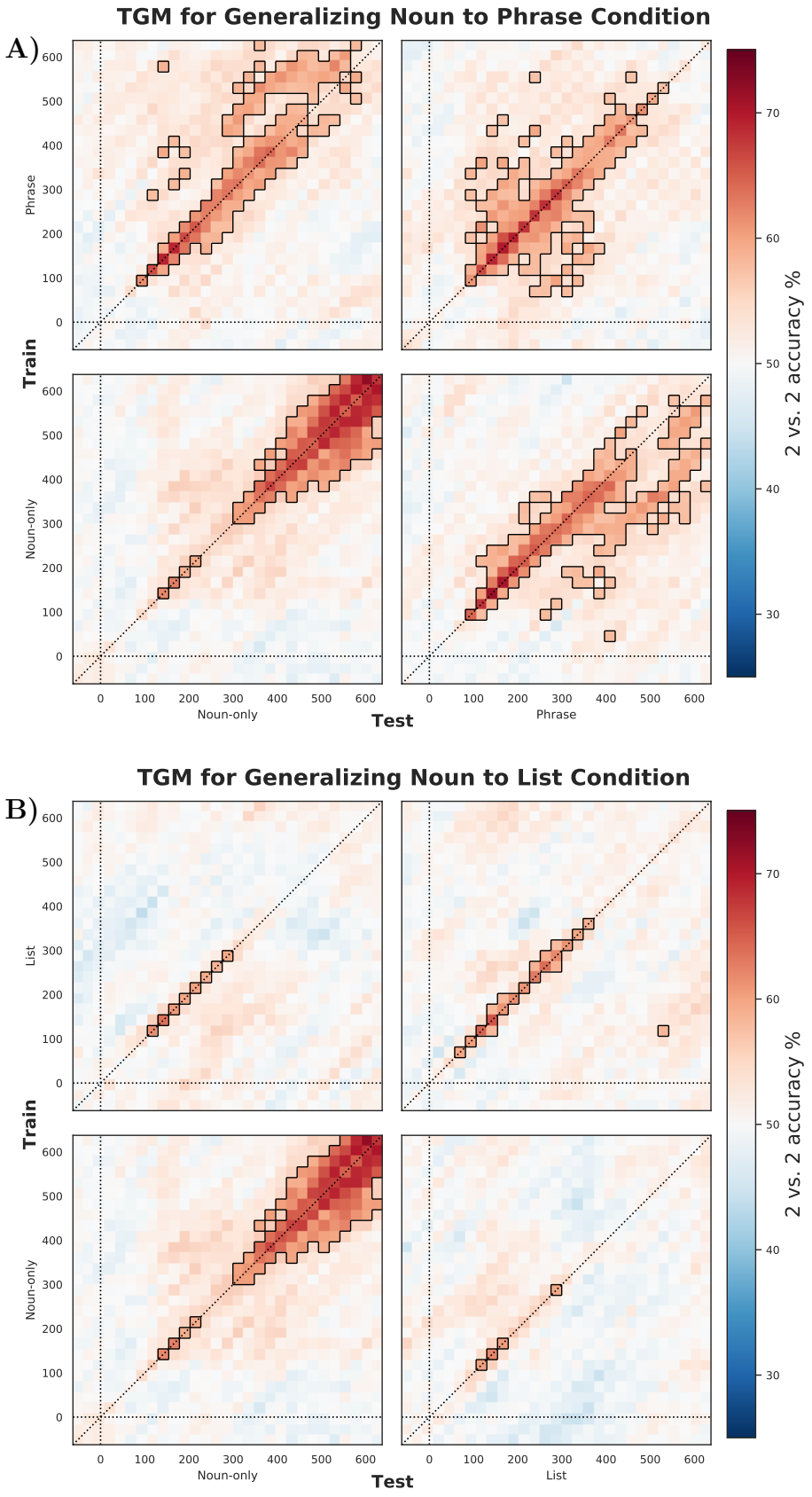

Figure 5.6: Temporal Generalization Matrices (TGMs) for decoding the noun semantics across A) noun-only and phrase conditions B) noun-only and list conditions. Each cell shows the 2 vs. 2 accuracy of a model trained and tested on the same or different time windows. All time windows are 100 ms long, starting from -100 ms with 25 ms sliding step. Significant cells are contoured ($p < .01$, FDR corrected over train and test time). Noun generalization gives higher 2 vs. 2 accuracy in (A) compared to (B).

condition temporally generalize to later time windows in the phrase condition. On the other hand, the phrase/noun-only TGM (Figure 5.6.A, top-left) has a significant patch above diagonal which demonstrates that model trained on the phrase condition temporally generalizes to earlier time windows in the noun-only conditions. These two TGMs (Figure 5.6.A, bottom-right and top-left) suggest a delayed repetition of noun representation in the phrase condition compared to the noun-only condition.

Next, Figure 5.6.B shows the TGMs for generalizing the noun within and across the noun-only and the list conditions. The noun-only/noun-only TGM (Figure 5.6.B, bottom-left) displays a late diagonal patch which relates to the noun utterance. There is also an early diagonal peak in this TGM which does not significantly generalize to any off-diagonal cells. We have a similar observation in the list/list TGM (Figure 5.6.B, top-right) where there are a few significant off-diagonal cells. In other words, the learned representations do not significantly generalize across time.

Furthermore, in the noun-only/list and the list/noun-only TGMs (Figure 5.6.B, bottom-right and top-left respectively), we did not find any significant off-diagonal cell. Comparing the TGMs in Figure 5.6.A and Figure 5.6.B, we observe that models trained or tested on the noun-only condition generalize better to the phrase condition rather than the list condition. In other words, the noun representation in preparation for production of an individual noun is more similar to that of an adjective-noun phrase than the list of an adjective followed by a noun.

5.4.2 TGMs for Inter-condition Decoding of the Adjective

We also tested the cross-condition generalization of the adjective with TGMs. Figure 5.7.A depicts the TGMs for generalizing the adjective representation between adjective-only and phrase conditions. All four TGMs in this figure contain a top-right significant patch corresponding to the models which are trained and tested close to the utterance, at ∼600 ms. These significant patches support our hypothesis of enhanced mental representation of the word close to its utterance.

Figure 5.7.B depicts the generalization of the adjective representation between the adjective-only and the list conditions. Recall that the adjective refers to the background colour in the list condition whereas it refers to the to the object's colour in the adjective-only condition. The background and the foreground colours are visually located in different parts of stimuli, which might affect the early perceptual processing of the adjective in the brain. In the list/list TGM (Figure 5.7.B, top-right),

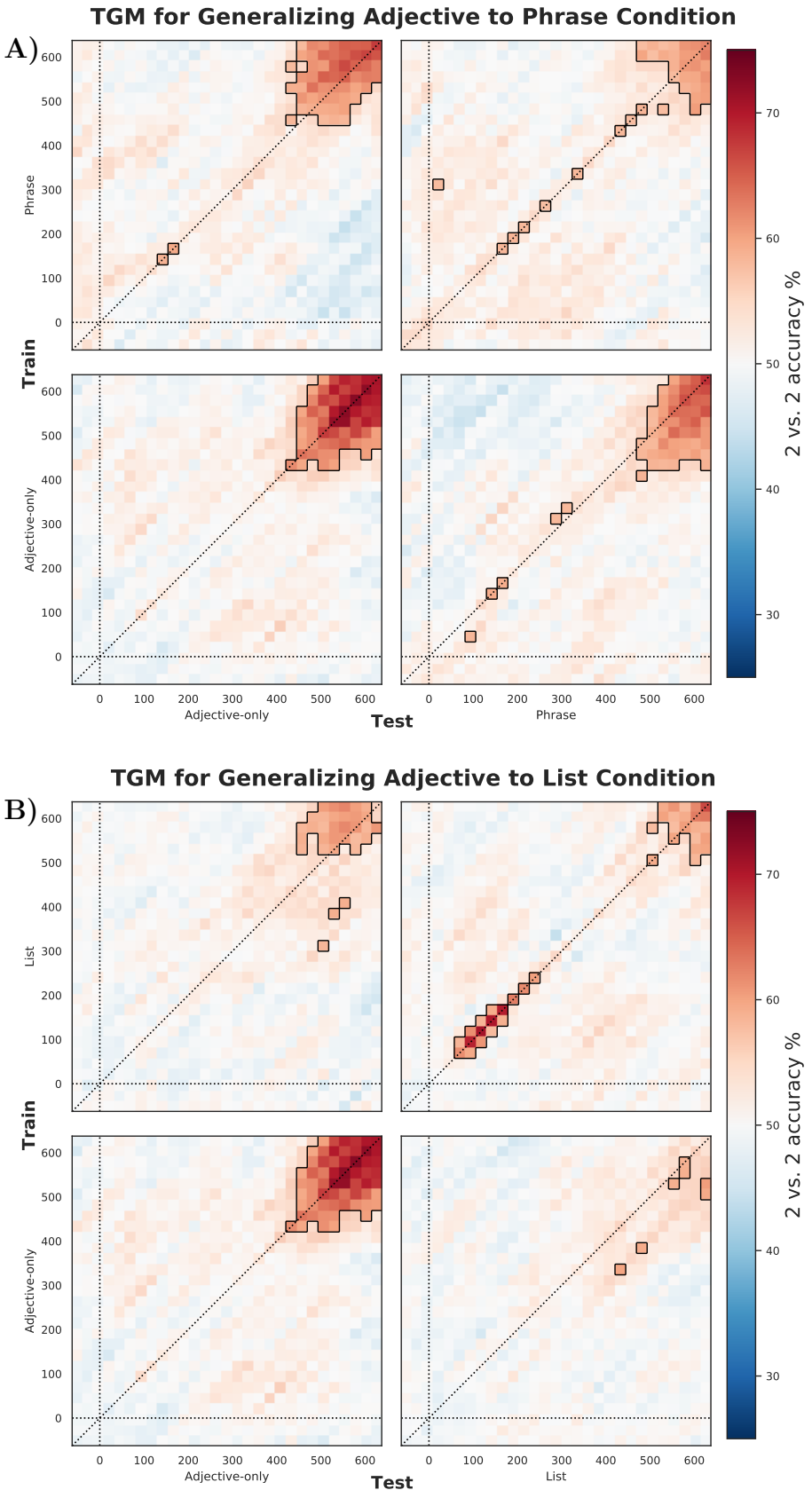

Figure 5.7: Temporal Generalization Matrices (TGMs) for decoding adjective semantics across A) adjective-only and phrase conditions B) adjective-only and list conditions. Each cell shows the 2 vs. 2 accuracy of a model trained and tested on the same or different time windows. All time windows are 100 ms long, starting from -100 ms with 25 ms sliding step. Significant cells are contoured ($p < .01$, FDR corrected over train and test time).

we observe an early and a late diagonal significant patches which do not generalize to each other (see section 5.4). The list/adjective-only TGM (Figure 5.7.B, top-left) has a late significant patch about ∼600 ms. This significant patch shows that the models trained on the list condition generalize to the adjective-only condition only in the later time interval close to the word's utterance while the earlier learned representation do not generalize. This pattern might suggest that the the adjective representation in the list condition during the early significant time interval has a visual representation while the later one potentially has a semantic or motor representation. However, this hypothesis needs further investigation.

5.5 Conclusion and Future Work

In this chapter, we analyzed semantic representation of the words in preparation of language production using the MEG data from a picture naming experiment designed and conducted by Blanco-Elorrieta et al., (2018) [11]. Our approach can significantly detect the semantic representation of the words which participants are about to say. Moreover, there was an enhanced semantic decodability close to the word utterance.

Comparing the temporal representation of the words in a non-compositional list condition with a compositional phrase condition, we found that composition enhances the noun decodability and is associated with a more durable noun representation in the brain. Further, TGMs analysis indicated a delayed repetition of noun representation in compositional adjective-noun phrases.

As a future extension of the current study, we suggest several directions to follow. In this thesis, we used semantic vectors for individual words. However, one could use phrase semantic representation such as additive and multiplicative models (Mitchell and Lapata, 2010) or ELMo representations (Peters et al., 2018) [23, 57].

We applied our decoding framework to the whole brain MEG sensor data. It would be valuable to explore smaller areas of cortex and identify what areas are contributing the most to the model performance and see how the word representation is processed in compositional related areas (i.e. LATL and vmPFC) identified in previous research [9, 10, 11].

In this experiment, the adjective set was limited to colour-describing adjectives. It would be interesting to explore other types of adjectives or more complex structures (i.e. idiomatic phrases or longer combination of words) during a language production task. Additionally, future experiments with reversed phrase structures such as

noun-adjective phrases in languages other than English can confirm our hypothesis of semantic enhancement close to the word's articulation.

Chapter 6

Conclusions

While recent studies have analyzed the mental representation of semantics during comprehension of individual words or adjective-noun phrases [3, 4, 5], the representation of words during utterance planning has been less explored. In this thesis, we investigated the semantic representation of words prior to their utterance. We also explored how this representation is changed when the words are combined in a compositional adjective-noun phrase.

Through previous chapters, we introduced some of the common distributional semantic models (DSM) and their application in the study of semantic representation in the human brain. Our data came from an MEG experiment designed and conducted by Blanco-Elorrieta et al., (2018) in which participants perform picture naming tasks with varying compositional requirements [11]. Based on the MEG recordings, we predicted the semantic properties of the words before their utterance, using existing brain decoding methods and DSMs. This methodology allows us to not only analyze the semantic representation of the words throughout time but also generalize the learned representations between the tasks.

Comparing the performance of our models trained on MEG recordings of compositional and non-compositional tasks, we provided evidence implying that:

- There is enough information in MEG data to decode the semantic representation of the words that the participant is about to say.
- Semantic decodability of a word improves close to its utterance.
- A compositional task enhances the decodability of the noun semantics, compared to a non-compositional task.

- A compositional task is associated with maintaining the noun representation in the brain for longer, compared to a non-compositional task.

While this thesis explores the compositional processes in the brain, it is still limited to simple adjective-noun phrases. Apart from that, the experimental design confined our analysis to colour-describing adjectives and object-describing nouns. However, there are many other compositional structures in human language which their underlying mental mechanisms remain to be explored.

Our research showed that existing brain decoding methods can be used to study and decode word semantics during speech planning. These methods can go beyond conventional event-related activity comparisons and explore the mental processes involved in language production tasks. In addition, the generalization of these machine learning models across time and experimental conditions can provide further evidence on the similarity of mental processes. The analytical framework used in this thesis could be applied to many other language-related experiments to further explain semantic composition in the brain.

Bibliography

[1] T. M. Mitchell, S. V. Shinkareva, A. Carlson, K. M. Chang, V. L. Malave, R. A. Mason, and M. A. Just, "Predicting human brain activity associated with the meanings of nouns," *Science (New York, NY)*, vol. 320, no. 5880, pp. 1191–1195, 2008. [Online]. Available: http://www.sciencemag.org/content/320/5880/1191.short{%}5Cnpapers://09b5ccc1-f521-4f79-8482-b284612215a7/Paper/p15147

[2] K. Chang, V. L. Cherkassky, M. Tom, and M. A. Just, "Quantitative modeling of the neural representation of adjective-noun phrases to account for fmri activation," 01 2009, pp. 638–646.

[3] B. Murphy, M. Poesio, F. Bovolo, L. Bruzzone, M. Dalponte, and H. Lakany, "EEG decoding of semantic category reveals distributed representations for single concepts," *Brain and Language*, vol. 117, no. 1, pp. 12–22, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.bandl.2010.09.013

[4] G. Sudre, D. Pomerleau, M. Palatucci, L. Wehbe, A. Fyshe, R. Salmelin, and T. Mitchell, "Tracking neural coding of perceptual and semantic features of concrete nouns," *NeuroImage*, vol. 62, 2012.

[5] A. Fyshe, G. Sudre, L. Wehbe, N. Rafidi, and T. M. Mitchell, "The Semantics of Adjective Noun Phrases in the Human Brain," *bioRxiv*, 2016.

[6] D. K. Bemis and L. Pylkkänen, "Simple composition: A magnetoencephalography investigation into the comprehension of minimal linguistic phrases," *Journal of Neuroscience*, vol. 31, no. 8, pp. 2801–2814, 2011.

[7] D. Bemis and L. Pylkknen, "Basic linguistic composition recruits the left anterior temporal lobe and left angular gyrus during both listening and reading," *Cerebral Cortex*, vol. 23, no. 8, pp. 1859–1873, 2013. [Online]. Available: http://dx.doi.org/10.1093/cercor/bhs170

[8] D. K. Bemis and L. Pylkkänen, "Flexible Composition: MEG Evidence for the Deployment of Basic Combinatorial Linguistic Mechanisms in Response to Task Demands," *PLoS ONE*, vol. 8, no. 9, 2013.

[9] L. Pylkkänen, D. K. Bemis, and E. B. Elorrieta, "Building phrases in language production: An MEG study of simple composition," *Cognition*, vol. 133, no. 2, pp. 371–384, 2014.

[10] P. Del Prato and L. Pylkkänen, "MEG evidence for conceptual combination but not numeral quantification in the left anterior temporal lobe during language production," *Frontiers in psychology*, vol. 5, p. 524, 2014.

[11] E. Blanco-Elorrieta, I. Kastner, K. Emmorey, and L. Pylkkänen, "Shared neural correlates for building phrases in signed and spoken language," *Scientific Reports*, vol. 8, no. 1, pp. 1–10, 2018.

[12] T. Mikolov, W.-T. Yih, and G. Zweig, "Linguistic Regularities in Continuous Space Word Representations," pp. 746–751, 2013.

[13] A. G. Huth, W. A. D. Heer, T. L. Griffiths, F. E. Theunissen, and L. Jack, "Natural speech reveals the semantic maps that tile human cerebral cortex," *Nature*, vol. 532, no. 7600, pp. 453–458, 2016. [Online]. Available: http://dx.doi.org/10.1038/nature17637

[14] H. Rubenstein and J. B. Goodenough, "Contextual correlates of synonymy," *Commun. ACM*, vol. 8, no. 10, pp. 627–633, Oct. 1965. [Online]. Available: http://doi.acm.org/10.1145/365628.365657

[15] J. R. Firth, "A synopsis of linguistic theory, 1930-1955," *Studies in linguistic analysis*, 1957.

[16] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.

[17] T. K. Landauer and S. T. Dumais, "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge." *Psychological review*, vol. 104, no. 2, p. 211, 1997.

[18] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[19] H. Schütze, "Dimensions of meaning," in *Proceedings of the 1992 ACM/IEEE Conference on Supercomputing*, ser. Supercomputing '92. Los Alamitos, CA, USA: IEEE Computer Society Press, 1992, pp. 787–796. [Online]. Available: http://dl.acm.org/citation.cfm?id=147877.148132

[20] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *Journal of machine learning research*, vol. 3, no. Feb, pp. 1137–1155, 2003.

[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Effecient Estimation of Word Representations in Vector Space," *CoRR*, pp. 1–12, 2013. [Online]. Available: http://arxiv.org/pdf/1301.3781v3.pdf

[22] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[23] J. Mitchell and M. Lapata, "Composition in distributional models of semantics." *Cognitive science*, vol. 34, no. 8, pp. 1388–1429, 2010.

[24] M. Baroni and R. Zamparelli, "Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2010, pp. 1183–1193.

[25] R. Socher, J. Bauer, C. D. Manning *et al.*, "Parsing with compositional vector grammars," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2013, pp. 455–465.

[26] K. M. Hermann and P. Blunsom, "The role of syntax in vector space models of compositional semantics," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2013, pp. 894–904.

[27] M. Yu and M. Dredze, "Learning composition models for phrase embeddings," *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 227–242, 2015.

[28] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin, "Placing search in context: The concept revisited," in *Proceedings of the 10th international conference on World Wide Web*. ACM, 2001, pp. 406–414.

[29] F. Hill, R. Reichart, and A. Korhonen, "Simlex-999: Evaluating semantic models with (genuine) similarity estimation," *Computational Linguistics*, vol. 41, no. 4, pp. 665–695, 2015.

[30] M. Baroni and A. Lenci, "How we BLESSed distributional semantic evaluation," in *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*. Association for Computational Linguistics, 2011, pp. 1–10.

[31] M. Baroni, G. Dinu, and G. Kruszewski, "Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2014, pp. 238–247.

[32] M. Faruqui, Y. Tsvetkov, P. Rastogi, and C. Dyer, "Problems with evaluation of word embeddings using word similarity tasks," *arXiv preprint arXiv:1605.02276*, 2016.

[33] M. Batchkarov, T. Kober, J. Reffin, J. Weeds, and D. Weir, "A critique of word similarity as a method for evaluating distributional semantic models," 2016.

[34] B. Chiu, A. Korhonen, and S. Pyysalo, "Intrinsic evaluation of word vectors fails to predict extrinsic performance," in *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, 2016, pp. 1–6.

[35] H. Xu, B. Murphy, and A. Fyshe, "Brainbench: A brain-image test suite for distributional semantic models," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 2017–2021.

[36] D. Dharmaretnam, "A Study of Semantics Across Different Representations of Language," Ph.D. dissertation, 2018.

[37] S. Abnar, R. Ahmed, M. Mijnheer, and W. Zuidema, "Experiential, distributional and dependency-based word embeddings have complementary roles in decoding brain activity," in *Proceedings of the 8th Workshop on*

Cognitive Modeling and Computational Linguistics (CMCL 2018). Association for Computational Linguistics, 2018, pp. 57–66. [Online]. Available: http://aclweb.org/anthology/W18-0107

[38] L. F. Nicolas-Alonso and J. Gomez-Gil, "Brain computer interfaces, a review," *Sensors*, vol. 12, no. 2, pp. 1211–1279, 2012.

[39] A. M. Glenberg, "How reading comprehension is embodied and why that matters," *International Electronic Journal of Elementary Education*, vol. 4, no. 1, pp. 5–18, 2017.

[40] L. W. Barsalou and K. Wiemer-Hastings, "Situating abstract concepts," *Grounding cognition: The role of perception and action in memory, language, and thought*, pp. 129–163, 2005.

[41] F. Pulvermüller, "Brain mechanisms linking language and action," *Nature Reviews Neuroscience*, vol. 6, no. 7, p. 576, 2005.

[42] T. M. Mitchell, R. Hutchinson, R. S. Niculescu, F. Pereira, X. Wang, M. Just, and S. Newman, "Learning to decode cognitive states from brain images," *Machine learning*, vol. 57, no. 1-2, pp. 145–175, 2004.

[43] T. M. Mitchell, S. V. Shinkareva, A. Carlson, K.-M. Chang, V. L. Malave, R. A. Mason, and M. A. Just, "Predicting human brain activity associated with the meanings of nouns," *science*, vol. 320, no. 5880, pp. 1191–1195, 2008.

[44] D. K. Bemis and L. Pylkkanen, "Combination across domains: an MEG investigation into the relationship between mathematical, pictorial, and linguistic processing," *Frontiers in psychology*, vol. 3, p. 583, 2013.

[45] G. Flick, Y. Oseki, A. R. Kaczmarek, M. Al Kaabi, A. Marantz, and L. Pylkkänen, "Building words and phrases in the left temporal lobe," *Cortex*, 2018.

[46] T. Leffel, M. Lauter, M. Westerlund, and L. Pylkkänen, "Restrictive vs. nonrestrictive composition: a magnetoencephalography study," *Language, cognition and neuroscience*, vol. 29, no. 10, pp. 1191–1204, 2014.

[47] J. Mitchell and M. Lapata, "Vector-based Models of Semantic Composition." *Acl*, vol. 8, no. June, pp. 236–244,

2008. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.164.9603{&}rep=rep1{&}type=pdf{%}5Cnhttp://homepages.inf.ed.ac.uk/s0453356/composition.pdf

[48] A. Fyshe, P. Talukdar, B. Murphy, and T. Mitchell, "Documents and Dependencies : an Exploration of Vector Space Models for Semantic Composition," *Conll*, pp. 84–93, 2013.

[49] S. G. Baron and D. Osherson, "Evidence for conceptual combination in the left anterior temporal lobe," *Neuroimage*, vol. 55, no. 4, pp. 1847–1852, 2011.

[50] S. L. Kivisaari, M. van Vliet, A. Hulten, T. Lindh-Knuutila, A. Faisal, and R. Salmelin, "Reconstructing meaning from bits of information," *bioRxiv*, p. 401380, 2018.

[51] A. J. Anderson, D. Kiela, S. Clark, and M. Poesio, "Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns," *Transactions of the Association of Computational Linguistics*, vol. 5, no. 1, pp. 17–30, 2017.

[52] Y.-P. Ruan, Z.-H. Ling, and Y. Hu, "Exploring semantic representation in brain activity using word embeddings," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 669–679.

[53] T. Hastie and R. Tibshirani, "Efficient quadratic regularization for expression arrays," *Biostatistics*, vol. 5, no. 3, pp. 329–340, 2004.

[54] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *Annals of statistics*, pp. 1165–1188, 2001.

[55] E. Maris and R. Oostenveld, "Nonparametric statistical testing of EEG-and MEG-data," *Journal of neuroscience methods*, vol. 164, no. 1, pp. 177–190, 2007.

[56] J. R. King and S. Dehaene, "Characterizing the dynamics of mental representations: The temporal generalization method," *Trends in Cognitive Sciences*, vol. 18, no. 4, pp. 203–210, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.tics.2014.01.002

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