Corpora and Cognition: The Semantic Composition of Adjectives and Nouns in the Human Brain

Alona Fyshe

February 2015
CMU-ML-15-100

Machine Learning Department
School of Computer Science
Carnegie Mellon University

Thesis Committee:
Tom Mitchell, Chair
Marcel Just
Byron Yu
Mirella Lapata

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.

This research was sponsored by Natural Sciences and Engineering Research Council of Canada (PGSD), Keck Foundation: grant number DT123107; National Science Foundation: grant number IIS0835797; National Institutes of Health: grant numbers U54GM088491, R90DA022761, and R90DA023420; Air Force Research Laboratory: contract number FA865013C7360.

Keywords: language, brain imaging, machine learning, distributional semantics, semantic composition

For my mother, Linda Ambrose, who made me think anything was possible,
to my husband, Mark Holzer, who made the impossible possible,
and to Jane Juniper, my little forcing function.

In memory of my father, Jonathan Fyshe who had a lifelong interest in the brain.

Abstract

The action of reading, understanding and combining words to create meaningful phrases comes naturally to most people. Still, the processes that govern semantic composition in the human brain are not well understood. In this thesis, we explore semantics (word meaning) and semantic composition (combining the meaning of multiple words) using two data sources: a large text corpus, and brain recordings of people reading adjective noun phrases. We show that these two very different data sources are both consistent, in that they contain overlapping information, and are complementary, in that they contain non-overlapping, but still congruent information. These disparate data sources can be used together to further the study of semantics and semantic composition as grounded in the brain, or more abstractly as represented in patterns of word usage in corpora.

This thesis is supported by three experiments. Firstly, we extend a matrix factorization algorithm to learn an interpretable semantic space that respects the composition of words into noun phrases. We use the interpretability of the model to explore semantic composition as captured in the statistics of word usage in a large text corpus. Secondly, we build a joint model of semantics in corpora and the brain, which fuses brain imaging data with corpus data into one model of semantics. When compared to models that use only a single data source, we find that this joint model excels at a variety of tasks, from matching human judgements of semantics to predicting words from brain activity.

Thirdly, we explore semantic composition in the brain through a new brain image dataset, collected with Magnetoencephalography while subjects read adjective-noun phrases. We learn several functions of the brain data that are capable of predicting semantic properties of the adjective, noun, and phrase. From the performance of these functions, we build a theory for the semantic composition of adjective noun phrases in the brain. This thesis asks a fundamentally different question than those asked in previous studies of adjective noun composition: where in the brain and when in time is *phrasal* meaning located? The answer to this question paints a unique picture of composition in the brain that is congruent with previous findings, but also sheds new light onto the neural processes governing semantic composition.

Together, these contributions show that brain imaging data and corpus data can be used in concert to build better models of semantics. These more successful models provide a new understanding semantic composition, both in the brain and in a more abstract sense. Furthermore, this thesis demonstrates how machine learning techniques can be used to analyze and understand complicated data, like the neural activity captured in brain images.

Acknowledgments

There are many people to thank for their support during my PhD, but Tom Mitchell certainly stands alone in his impact on my scholarly life. Tom's fresh thinking and positive attitude makes doing research with him a pleasure, and he contributed greatly to a graduate student experience that I will remember fondly.

I must also thank my lab mates, especially Gustavo Sudre who helped me to collect the MEG data were used in the experiments of this thesis. Gus was (and is) patient and incredibly generous with his time, and even took a day off work to help us finish source localizing our MEG data. Thank you, Gus!

I wold also like to thank Leila Wehbe and Nicole Rafidi for the many productive conversations and for their help getting me unstuck from research ruts. Thank you to Erika Laing for her help collecting data and for her language expertise. Thank you to Dan Schwartz and Dan Howarth for good conversations and technical support along the way.

To Kamal Nigam and Tom Murphy IV who encouraged me to apply to CMU, though I was doubtful I would be accepted. They had a tremendous impact on me during my time at Google, and from them I learned new ways of thinkng and how to best contribute to a large software project.

Thank you to my family. To my mother, who was the best female role model a girl could ask for: courageous, tenacious and a leader in her field. I never wondered if a woman could be a scientist (or a *computer* scientist!) because I was raised without such assumed gender barriers. It's amazing what you can do when the people around you think you're capable of anything. To my brother for being supportive and interested in my research, and for being a medical professional who is willing to teach me about the amazing things he sees at work.

And last but never least, to my husband Mark, whom I am so lucky to have in my life. When I was offered a position at Google Pittsburgh, Mark moved with me without ever having visited the city. And then when I left my (well paying) job to return to graduate school, he was supportive and proud. When I struggled during the first years of school, Mark was there to pick me up and encourage me to keep trying. Towards the end of the program, when I was working every waking hour to finish my thesis, he cooked me homemade meals and kept me in clean clothes when otherwise I would certainly have had neither. Mark, you are my best friend, my confidant, my support, and there aren't words to express my gratitude for having you in my life. I love you more every day.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The action of combining words to create higher order meaning comes naturally to most people. Still, the brain processes that govern the combining of linguistic units are not well understood. This thesis explores the mechanisms by which the human brain retrieves the meaning of adjectives (e.g. tasty) and nouns (e.g. tomato), and combines them to form phrases with new and altered semantic meanings. To this end, we utilize both written language resources (corpora) and recordings of the human brain (cognition).

The fields of Computational Linguistics and Psycholinguistics have studied word meaning (semantics) and semantic composition (the process of combining smaller linguistic units to make more complex meaning) for decades. Computational Linguists have been blessed with an abundance of text data freely available for download on the Internet. This has allowed researchers to build models of semantics using patterns of word usage, models that are surprisingly consistent with human judgements of word meaning (Lund and Burgess, 1996b; Landauer and Dumais, 1997b; Sahlgren, 2006a). Psycholinguists use the power of brain imaging, which allows them to peer into the brain as a subject performs a neural semantic analysis of words during reading. This has allowed us to better understand which areas of the brain are implicated in the semantic and syntactic tasks involved in comprehending language. Though the fields of Computational Linguistics and Psycholinguistics have begun to work in collaboration, they tend to be largely separate with very different ways of approaching the study of semantics and semantic composition.

This fracture of Computational Linguistics and Psycholinguistics is lamentable because the two fields study the same phenomena from two different vantage points. Computational linguists often study language using large text corpora, which are the output of many human brains communicating via language. Psycholinguists, on the other hand, measure correlates of one brain comprehending language, either through behavioral measurements (e.g. question answering, eye tracking or response time measurements) or through the measurement of brain activity. Though the two fields approach the problem from two very different angles, they both seek to uncover the organization and structure of language, which has its roots in the neural substrate of the brain.

Corpus data and brain imaging data have distinct advantages and disadvantages. For example, brain imaging data is expensive to collect but is a more direct measurement of neural semantic representations. Corpus data, on the other hand, has the advantage of being cheap and plentiful, but the disadvantage of being noisy and suffering from linguistic artifacts like polysemy. Algo-

rithms to combine corpus data and brain imaging data, or that leverage one data source to study the other, could use the strengths of one data source to compensate for the weaknesses in the other.

1.1 Thesis Statement and Contributions

The central thesis of this work is:

> *Corpora and brain imaging data sources, which are both consistent and complementary, can be used together to further the study of semantics and semantic composition. We can use machine learning algorithms to learn functions of each data source, or functions over both data sources simultaneously. The output of these functions, and the functions themselves, can be used to study semantics as grounded in the brain, or more abstractly as represented in patterns of word usage in corpora.*

To support this thesis, we bring together advances from the fields of Computational Linguistics and Psycholinguistics to further the study of semantics and semantic composition. For this thesis, Machine Learning algorithms will be pivotal, as they will help us to analyze corpora compiled from millions of web pages, and the complex images produced by brain imaging technologies like functional Magnetic Resonance Imaging (fMRI) and Magnetoencephalography (MEG). We use machine learning to build several functions that take as input corpus data, brain data or some combination of the two. We will use the output of these learned functions, or the functions themselves, as a tool to better understand the input data source.

One of the main focuses of this thesis is semantic composition, specifically adjective-noun composition. Adjective-noun phrases represent a very basic form of semantic composition, but they have several advantageous properties. Firstly, the neural representation of nouns in isolation has been studied extensively (Mitchell et al., 2008; Palatucci et al., 2009; Just et al., 2010; Chang et al., 2011; Sudre et al., 2012). Secondly, and in concert with the first advantage, nouns and adjective-noun phrases are both noun phrases, so they live in the same semantic space. This allows us to leverage previous work on the representation of nouns in the brain to study noun phrase composition in the brain. Thirdly, adjective-noun composition has the advantage of being a recent topic of interest amongst Computational Linguists (Mitchell and Lapata, 2010; Turney, 2012) as well as psycholinguists (Bemis and Pylkkänen, 2011; Baron, 2012). These assets make adjective-noun phrases an attractive option for studying semantic composition.

This thesis has three main contributions, covered in three chapters:

1. **Chapter 3: An Interpretable Model of Semantic Composition** Previously proposed models of semantic composition use latent representations that are difficult to interpret and are often created with methods that do not directly model composition. In this chapter we learn latent semantic spaces that represent composition in an interpretable way and outperform several previously proposed methods. We illustrate how interpretability allows us to explore performance on several compositional tasks, paving the way for the improvement of subsequent compositional models.

2. **Chapter 4: A Joint Model of Semantics in Corpora and the Brain** Typically, semantic models are built using large corpora, and these models have been used to study semantics in the brain. This chapter uses machine learning to fuse brain imaging data with corpus data into one joint model of semantics. When compared to models that use only a single data source, this joint model excels at a variety of tasks, which shows that corpus and brain imaging data are not only consistent, but complementary.
3. **Chapter 5: Semantic Composition in the Brain** Very little is known about the way the human brain represents composed phrasal meaning. Here we introduce a new brain imaging dataset collected during adjective-noun phrase reading, and use it to conduct the first study of phrasal semantics in the brain. We build several functions over the brain data to predict properties of the adjective, noun, and phrase. We use the performance of these functions to infer which areas of the brain are involved at which times as adjective noun phrases are read, composed and understood.

As a precursor to these three content chapters, Chapter 2 covers the most relevant related work from computational linguistic and psycholinguistic studies of semantics. Chapter 6 concludes the thesis and brings together the findings from each chapter.

The contributions of this thesis show that brain imaging data and corpus data can be used in concert to build better semantic models of word and phrasal meaning. Our improved models allow us to deepen our understanding of semantics and semantic composition, both in the brain and in a more abstract sense. Furthermore, we demonstrate how Machine Learning techniques can be used to analyze and understand the complicated neural activity captured in brain images.

Chapter 2

Related Work

This thesis covers two broad research areas - the study of language using corpora (text collections) and the study of language in the brain via brain imaging. The two fields aim to measure language and word meaning, but approach it in two very different ways. Considering both areas of research simultaneously may help us build a better understanding of semantics and language. Therefore, this chapter gives the relevant history of study of semantics in text, as well as language in the brain.

2.1 Semantics in Text

The computational study of semantics has been greatly influenced by the idea that word meaning can be inferred by the context surrounding a given word, averaged over many examples of the word's usage. For example, we might see the word *ball* with verbs like *kick*, *throw*, *catch*, with adjectives like *bouncy* or with nouns like *save* and *goal*. This observation prompted linguist John Firth to state "You shall know a word by the company it keeps". Context cues give us an idea of how a ball is used and thus what the meaning of ball might be.

This inference of word meaning is something adults do naturally when they encounter an out of vocabulary word in text. Often, we can infer an unfamiliar word's meaning by the theme of the passage and the words near the unknown word. The idea that context equals semantics drives much of the work on models of semantics as derived from large bodies of text.

2.1.1 Distributional Semantics

Distributional Semantics leverages the idea that word usage implies word meaning. Large collections of text, often gathered from the Internet, are used to compile statistics about word usage, which can then be used to create a model of word meaning. In Vector Space Models of semantics (VSMs), each word is assigned a vector, and the elements of the vector correspond to corpus statistics collected about the word. These statistics can include word-document co-occurrence (e.g. the word *ball* was seen 10 times in document 400), word-word co-occurrence within some window (e.g. the word *ball* and the word *goal* appear together within a 5 word window 20 times in the corpus), or word-word-dependency triples (e.g. the word *ball* was the subject of the verb

kick 19 times in the corpus). Each one of these statistics could become an element in *ball*'s word vector representation. Especially when word-word-dependency triples are used, the number of elements in each word's vector can become very large, and often very sparse (e.g. the word lettuce may never occur near ball, nor even in the same documents as the word ball). For this reason, compression of the word vectors is often performed with a technique such as singular value decomposition (SVD). This compression of word statistics is the basis of latent semantic analysis (LSA) (Landauer and Dumais, 1997a), one of the seminal works in distributional semantics.

Further research in distributional semantics determined the types of corpus-derived statistics that were most useful for particular semantic tasks. These studies tend to compare features derived from global corpus co-occurrence patterns (e.g. how often a word appears in each document), or local corpus co-occurrence patterns patterns (e.g. how often two words appear together in the same sentence, or are linked together in dependency parsed sentences). These two feature types represent different aspects of word meaning (Murphy et al., 2012c; Turney, 2012; Fyshe et al., 2013), and can be compared with the paradigmatic (words are substitutable for each other) syntagmatic (words are commonly seen together) distinction ofSahlgren (2006b). Global patterns give a more *topic-based* meaning. For example *athlete* might appear in documents also containing *field* and *score*. Certain local patterns give a more *type-based* meaning. For example, the noun *athlete* might be modified by the adjective *talented*, or be the subject of *scored*, as would the substitutable words such as *player* or *teammate*. Global patterns have been used in Latent Semantic Analysis and LDA Topic models (Blei et al., 2003). Local patterns based on word co-occurrence in a fixed width window were used in Hyperspace Analogue to Language (Lund and Burgess, 1996a). Subsequent models added increasing linguistic sophistication, up to full syntactic and dependency parses (Lin, 1998; Padó and Lapata, 2007; Baroni and Lenci, 2010).

Recently, Baroni et al. (2014) reported that for several kinds of semantic tasks, including semantic relatedness and analogy tasks, vectors built by collecting statistics about word-word collocation were outperformed by a neural language model. Baroni compared the typical count-based word vectors to vectors produced by a multi-layer neural network (similar to Socher et al. (2012)) trained to predict a word in a piece of text, given the context (Mikolov et al., 2013),. The output of these models is a vector used to make the context prediction for a given word. Subsequently, some forms of neural language models have been shown to be equivalent to matrix factorization applied to windowed, normalized word co-occurrence counts (Levy and Goldberg, 2014).

2.1.2 Composition for Distributional Semantics

Distributional semantics have also been used to model semantic composition: the process by which meaning is formed or altered by the combination of words. Generally, the purpose of these semantic composition techniques is to approximate the vector for a phrase p (and thus the phrase's meaning) by applying some function f to the phrase's constituent words. So, if phrase p is composed of words $w^{(1)}, w^{(2)}$ we wish to approximate the function:

$$f(\vec{w}^{(1)}, \vec{w}^{(2)}) = \vec{p} \tag{2.1}$$

where $\vec{w}^{(1)}, \vec{w}^{(2)}$ and $\vec{p}$ are the VSM vectors for each of the word/phrase units.

Mitchell and Lapata (2010) explored several methods of combining adjective ($\vec{w}^{(1)}$) and noun ($\vec{w}^{(2)}$) vectors to estimate phrase ($\vec{p}$) vectors, and compared the similarity judgements of humans to the similarity of their predicted phrase vectors. They found that for adjective-noun phrases, type-based models outperformed Latent Dirichlet Allocation (LDA) topic models. Mitchell and Lapata explore several composition functions including weighted addition, multiplication, and dilation Weighted addition of two semantic vectors $\vec{w}^{(1)}$ and $\vec{w}^{(2)}$ is defined as:

$$\hat{p}_i = \alpha w_i^{(1)} + \beta w_i^{(2)} \tag{2.2}$$

where α and β are parameters to be learned, and i denotes the ith element of the vector. Multiplication of two vectors is simply the element-wise product of the vectors:

$$\hat{p}_i = w_i^{(1)} * w_i^{(2)} \tag{2.3}$$

Dilation of two semantic vectors, an adjective ($\vec{w}^{(1)}$) and a noun ($\vec{w}^{(2)}$) involves breaking the noun into a component parallel to the adjective ($\vec{x}$) and a component perpendicular to the adjective ($\vec{y}$):

$$\vec{x} = \frac{\vec{w}^{(1)} \cdot \vec{w}^{(2)}}{\vec{w}^{(1)} \cdot \vec{w}^{(1)}} \cdot \vec{w}^{(1)} \tag{2.4}$$

$$\vec{y} = \vec{w}^{(2)} - \frac{\vec{w}^{(1)} \cdot \vec{w}^{(2)}}{\vec{w}^{(1)} \cdot \vec{w}^{(1)}} \cdot \vec{w}^{(1)} \tag{2.5}$$

$$\tag{2.6}$$

where $\cdot$ is the vector dot product. Then we compute dilated composition by enhancing the component parallel to the adjective (x) by multiplying it by a scalar (γ):

$$\hat{\vec{p}} = \gamma\vec{x} + \vec{y} \tag{2.7}$$

For type-based models, multiplication of the vectors performed the best, followed by weighted addition and dilation. Two other comparisons of vector-space representations found that the best performance for adjective-noun composition used point-wise multiplication and a model based on type-based word co-occurrence patterns (Blacoe and Lapata, 2012; Dinu et al., 2013).

Baroni and Zamparelli (2010) extended the typical vector representation of words. Their model used matrices to represent adjectives, while nouns were represented with column vectors. The vectors for nouns and adjective-noun phrases were derived from local word co-occurrence statistics. The matrix to represent the adjective was estimated with partial least squares regression where the product of the learned adjective matrix ($W^{(1)}$) and the observed noun vector ($\vec{w}^{(2)}$) should equal the observed adjective-noun phrase vector ($\vec{p}$).

$$W^{(1)} * \vec{w}^{(2)} = \vec{p} \tag{2.8}$$

Socher et al. (2012) also extended word representations beyond simple vectors. Their model assigns a vector and a matrix to each word. The vector and matrix are composed via the nonlinear function $tanh$ to create phrase representations, which consist of another vector/matrix pair.

This process can proceed recursively, following the parse tree for a sentence to produce a composite sentential meaning. Under this formulation, operator words that do not have attributional semantic content (like "better") will build up the matrix component, and the vector component will be driven to zero. Conversely, if a word is more contentful than operational (like "purple"), the matrix will be close to identity, and the vector will adapt to represent the word's semantics.

Other general semantic composition frameworks have been suggested, e.g. Sadrzadeh and Grefenstette (2011) who focus on the operational nature of composition, rather than the representations that are supplied to the framework. This idea has been extended to a more general approach to semantic composition within sentences (Grefenstette et al., 2013; Krishnamurthy and Mitchell, 2013; Hermann et al., 2013) which can use vectors, matrices and/or tensors for composition. However, scaling up to tensors requires the estimation of many more parameters which can be hard to tune with limited data. Some methods for handling parameter learning in this expanded space have been developed (Socher et al., 2013).

Turney (2012) explored of the impact of domain- and function-specific vector space models, analogous to the topic and type based corpus statistics mentioned previously in Section 2.1.1. In Turney's work, domain-specific information was represented by noun token co-occurrence statistics within a local window, and functional roles were represented by generalized token/part-of-speech co-occurrence patterns with verbs. Turney explored several compositional relations, including adjective-noun and noun-noun composition. Each word vector has a domain and function specific part. Then, to determine if a a phrase ab (composed of words a and b) is synonymous with word c, Turney used a hand-crafted comparison function $sim_c(ab, c)$:

$$sim_1(ab, c) = geo(sim_d(a, c), sim_d(b, c), sim_f(b, c)) \tag{2.9}$$

$$sim_c(ab, c) = \begin{cases} sim_1(ab, c) & \text{if } a \neq c \text{ and } b \neq c \\ 0 & \text{otherwise} \end{cases} \tag{2.10}$$

where $sim_d(x, y)$ is the similarity of the domain components of x and y, $sim_f(x, y)$ is the similarity of the function components of x and y, both calculated using cosine similarity. This comparison function is interesting because it is not symmetric. That is, the similarity of phrase ab to word c would be different than that of phrase ba to word c. Turney found this comparison function worked well for identifying single word synonyms for adjective noun and noun noun phrases.

2.2 Language in the Brain

2.2.1 Brain Imaging Modalities

Before brain imaging technologies were developed, the study of language in the brain used reaction times and eye tracking, and a considerable amount of progress was made with these simple measurements. More sophisticated brain imaging technologies have become very popular in recent decades, and have allowed researchers to explore the brain's activity during a variety of tasks.

The most common brain imaging technologies are Electroencephalography (EEG), Magnetoencephalography (MEG) and functional Magnetic Resonance Imaging (fMRI). Each technique

has its own unique advantages and disadvantages, and each measures brain activation in a different way.

Electroencephalography (EEG) is the oldest of the three brain imaging technologies discussed here. EEG measures the voltage fluctuations along the scalp that occur when many neurons fire in a coordinated fashion. EEG has the benefit of being able to record changes in voltage by the millisecond, making it one of the best brain recording modalities in terms of time resolution (similar resolution to MEG). EEG also has the benefit that it can be used in fairly uncontrolled settings like hospital rooms or offices, rather than the magnetically shielded rooms required by MEG or fMRI. The largest drawback of EEG is poor spatial resolution, which is caused by interference from the skull and scalp. This gives EEG spatial resolution on the order of 7 mm (Im et al., 2007), the worst amongst the three modalities described here.

MEG measures the magnetic field caused by many neurons firing in synchrony. An example MEG recording appears in Figure 2.3. MEG primarily measures the post-synaptic currents in apical dendrites that reside mostly in the sulci of the brain (Hansen et al., 2010). That is, MEG measures the currents caused by external neurons sending signals to (synapsing on) groups of neurons that lie parallel to the skull. Like EEG, MEG has time resolution on the order of ms, but magnetic fields do not suffer the same dampening by the skull and scalp as seen in EEG. For this reason the spatial resolution of MEG is better than EEG, as good as 2-3 mm (Hamalainen et al., 1993).

The imaging technique with the greatest spatial resolution is fMRI, which can achieve resolution as fine as 1mm. An example fMRI image appears in Figure 2.4. fMRI measures changes in blood oxygenation in response to increased neuronal activity, called the blood-oxygen-level dependent (BOLD) response. Because fMRI depends on the transport of oxygen via blood to the brain, its time constant is governed by the rate at which blood can replenish oxygen in the brain. Though fMRI can acquire images at the rate of about 1 image per second, the BOLD response can take several seconds to reach its peak after a stimulus is shown. Thus, amongst the three modalities discussed here, fMRI has the worst time resolution and the best spatial resolution.

2.2.2 Broca, Wernike, and the Dual Stream Hypothesis

Early studies of language in the brain began in the 1800s, when Paul Broca and Karl Wernike studied patients with brain injuries that affected their ability to communicate with language (Bear et al., 2007). Broca's studies of patients with Aphasia (partial or complete loss of language abilities as a result of brain injury) prompted him to conclude that language is controlled by only one hemisphere of the brain, almost always the left hemisphere. Broca's work also led him to identify a region of the brain in the posterior inferior left frontal gyrus which, when damaged, leads to non-fluent aphasia. Non-fluent aphasia, also called Broca's aphasia or expressive aphasia, is characterized by the inability to produce language, or by having great difficulty producing language. Often cognition is not impaired, and language can be understood, but the production of language is greatly hindered. The affected area of the posterior inferior left frontal gyrus has since been named Broca's area.

Wernike also found that lesions in the left hemisphere created language deficits. However, Wernike focused on an area of the brain in the posterior superior temporal gyrus, now called Wernike's area. Damage to this area results in a different type of aphasia, called fluent aphasia,

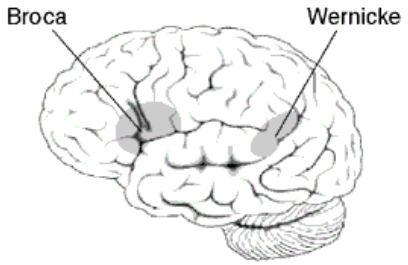


Figure 2.1: Broca's area and Wernike's area highlighted in the left hemisphere of the human brain. Image licensed under public domain via Wikimedia Commons.

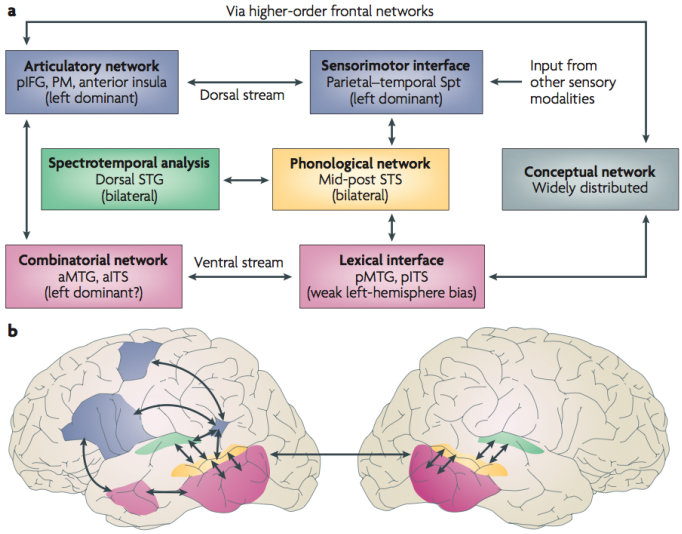


Figure 2.2: The Hickok and Poeppel (2007) model of language processing with dorsal and ventral streams inspired by vision research. **a.** Description of dorsal and ventral stream processing. **b.** Location of brain areas involved in dorsal and ventral streams. Image from Hickok and Poeppel (2007).

jargon aphasia or Wernike's aphasia. Fluent aphasia is characterized by the easy production of language that is mostly non-sensical or wandering. Intonation and speed of speech are usually normal, and if one ignores the content of the utterance, the speech can seem quite typical. In these patients, comprehension is difficult to assess, because responses are nonsensical. When instructions are given verbally, patients often have difficulty performing the requested actions, indicating that language understanding is also hindered by such brain injuries. These symptoms have lead to the theory that Wernike's area is involved in the mapping of language sounds or written words to semantic content. When Wernike's area is damaged, both written and vocal production (and comprehension) of meaningful language are negatively affected, but the physical act of producing written or spoken language is unhindered (though the produced language is illogical).

More recently, the brain areas theorized to be used for language processing have been extended. Hickok and Poeppel (2004, 2007) drew inspiration from vision research, which has broken visual processing into two streams: ventral (what) and dorsal (where). This theory is called the Dual Stream Hypothesis. The "what" stream handles identifying objects, whereas the "where" stream guides actions in response to stimuli and integrates visual and motor information (sometimes called the "how" stream). Hickok and Poeppel apply this two-stream hypothesis to language processing. The translation of the ventral stream is fairly straightforward, as the processing of semantics has been indicated in the temporal lobe of the brain (ventral). The dorsal stream is proposed to link the motor areas of the brain (including the articulatory network in the posterior inferior frontal gyrus) with auditory and sensorimotor areas of the brain. This new

model is consistent with the Broca and Wernike aphasias, which would result from damage to the ventral and dorsal streams of the Hickok and Poeppel model.

2.2.3 Composition in the Brain

Semantic composition is one part of a larger cognitive process termed semantic unification. Semantic unification includes not only composing the meaning of words in phrases, but also phrases in sentences, and sentences in larger thematic structures.

Semantic unification has been studied using EEG for decades. Early on, Kutas and Hillyard (1980) noted a more negative current in centro-parietal sensors due to a semantically mismatched sentence ending (e.g. *He spread the warm bread with socks.*). They named the phenomenon **N400**, and it has been widely studied since. Originally thought to be a reaction to semantically incompatible words, it has since been shown that the N400 can be evoked by sentences with semantically less predictable words. For example, "Jenny put the sweet in her (mouth/pocket) after the lesson" elicits an N400 for the word pocket. Though pocket is a semantically fine word choice, it is judged less probable than the alternative (mouth). It has also been shown that the N400 can appear before the incongruent word if the indefinite article (a/an) does not match the predicted noun. For example, an N400 will occur for the word *an* in the sentence"It was breezy, so the boy went to fly an kite" because the word kite is so strongly predicted and an is the wrong indefinite article.

In contrast to the N400, the P600 is characterized by a positive-going current that peaks around 600ms after stimulus onset, also in centro-parietal sensors. The P600 is closely tied to syntactic anomalies (whereas the N400 is associated with semantic mismatches), but has some interesting special cases. Typically, the P600 is present if the stimuli shows a syntactic violation (e.g. word order mistakes, plural verb disagreement, grammatical gender mismatch) (Kuperberg, 2007) . However, under certain circumstances a P600 can be evoked even when the syntax of a sentence is correct. For example, the sentence "Every morning the eggs would eat toast for breakfast." will induce a P600 for the underlined word "eat", though the sentence is syntactically sound, and elicits no N400, though the sentence is semantically incongruent. This phenomenon was called a "semantic illusion" because it fools the subject into thinking that the word is semantically sound due to their strong conceptual link (Hoeks et al., 2004).

It has been noted that many of the sentences that evoke this semantic P600 are mismatches in animacy; the verb requires an animate agent, but the agent supplied is inanimate. Three explanations for this behavior are supplied by Kuperberg (2007):

- Animacy is special, and is processed as a syntactic rather than semantic constraint.
- In the case of animacy mismatches, the cognitive process that combines words supplies thematic roles to a semantic cognitive process, where the verb is found to be incompatible with the noun. This triggers ongoing combinatorial analysis.
- The stream that handles combinatorial analysis recognizes some primitive semantic features (like animacy) and uses them to build the syntactic parse.

In addition, Sudre et al. (2012) found that animacy was the earliest semantic feature to be processed in the brain for concrete nouns, another piece of evidence that animacy is processed differently,

2.2.4 Proposed models of Semantic Unification

Kuperberg (2007) proposes a model for language processing based on the characteristics of stimuli that evoke the N400 and P600. Kuperberg postulates that there are at least two parallel processes involved:

- **Semantic Memory:** One stream responsible for the retrieval of semantic features, associations and semantic relations between words. N400 sensitivity
- **Combinatorial:** At least one stream (possibly multiple streams) responsible for the combination of words to create higher order meaning. P600 sensitivity

Under this model, the P600 is due to the continued analysis that takes place if the output of this combinatorial stream is incongruent with the output of the predictions of the semantic stream. The combinatorial stream processes two types of constraints

- **Morphosyntactic:** Attempts to resolve syntax errors.
- **Semantic-thematic:** can influence the N400 because it operates in parallel with the semantic memory stream. Processing of this constraint may continue after the N400 window if more combinatory analysis is needed.

An alternative model is proposed by Hagoort (2005), who describes a system for language in the brain drawing from fMRI evidence. The Hagoort MUC model consists of 3 functions:

Memory recalling the meaning of a word, lexical access. The temporal cortex and the inferior parietal cortex are involved in the memory process of Hagoort's model.

Unification integrating the retrieved meaning of a word with the meaning representation calculated with the context leading up to that word. This includes non-linguistic sources of meaning like gesture and gender of speaker. This processing resides in left inferior frontal cortex, including Broca's area (Brodmanns Area (BA) 44 and BA 45). BA 47 and BA 45 are involved in semantic unification, while syntactic unification is handled by BA 45 and BA 44 (Hagoort, 2014).

Control governs the actions required for language, like taking turns during a conversation. Control requires dorsolateral prefrontal cortex, anterior cingulate cortex (ACC) and the parts of parietal cortex that govern attention.

2.2.5 Semantics in the Brain

Semantics in the brain has historically been studied not by comparing the *magnitude* of activity between conditions, but rather by the information encoded in the neural activity. One can measure the information encoded in neural activity by training machine learning algorithms to predict some feature of the input stimuli. Machine learning algorithms do not require large differences in magnitude between conditions, but rather leverage patterns in the recordings of neural activity, which may involve differences in signal in both the positive and negative direction in different areas of the brain at different times. We will discuss Machine learning to recover the neural information encoding in greater detail in Chapter 5.

The study of semantics in the brain has often linked brain activation to linguistic measurements of semantics. Mitchell et al. (2008) showed that the fMRI activity of people reading 60 common concrete nouns could be modeled as the linear combination of features derived from

verb co-occurrence with the target word (essentially a simple VSM). Chang et al. (2011) extended this work to show that similar results could be obtained using feature norms (free-form naming of word characteristics), and Just et al. (2010) showed that the activity from noun reading could be tied to biologically relevant brain areas (e.g. manipulation-related words to motor cortex). Sudre et al. (2012) used Magnetoencephalography (MEG), the same 60 words of Mitchell et al., and behavioral data collected via Mechanical Turk[1] to explore the neural basis of semantic representation. Sudre et al. found that the semantic representation of a word unfolds over time, and that different semantic elements appear at different times in different parts of the brain. Murphy et al. (2012a) showed that a set of automatically derived corpus statistics (a VSM) could perform as well as the behavioral data from Sudre et al. (2012).

Another linguistic resource, WordNet, has also been used to study language in the brain. WordNet (Fellbaum, 1998) is a lexical database where English words and relationships between words are recorded (e.g. cat "is a kind of" feline). Words may be associated with groups of synonymous words called "synsets". Huth et al. (2012) annotated 2 hours of video with over 1700 WordNet categories. These annotations were then used to map semantic categories onto the brain via linear regression. The study confirmed much that was already known about semantics in the brain (e.g. face stimuli give strong reactions in the fusiform face area - FFA) but also showed the extremely distributed nature of semantics in the brain. For example, videos containing people show activation in FFA, but also in posterior Superior Temporal Sulcus (pSTS - associated with gaze following), in the Extrastriate Body Area (EBA - activated by stimuli containing body parts) as well as widespread activation in frontal and temporal regions. A brain-browsing interface has been supplied by the authors (`http://gallantlab.org/brainviewer/huthetal2012/`) which can be used to explore WordNet in cortical space.

Recently, MEG has ben used to study the effect of context on brain activation while subjects read a chapter from a story. Wehbe et al. (2014) used different linguistic techniques were used to represent semantics - Recurrent Neural Network Language Models (RNNLM) (Mikolov, 2012) and Neural Probabilistic Language Models (NPLM) (Vaswani et al., 2013). Each of these two models is a multi-layer neural network which represents the history of words encountered. In the case of RNNLM, an unlimited lexical history is available, constrained only by the size of the hidden layer in the network, whereas a 3- or 5-word history is used to train a NPLM. Both models are trained to predict the next word, given the word's previous context. Then a model was trained to predict story-reading MEG activity from the hidden, output or embedding layers of the neural networks. Wehbe et al. found that the hidden layer of a RNNLM performed best, followed by the hidden layer of a NPLM given 5 words of context. Context vectors were most useful for predicting brain activity 250ms after the onset of a word, perhaps reflecting the process of combining a new word with the current semantic state. Thus, Wehbe et al. show that story context can be used to differentiate brain states, and that some amount of the brain activation is correlated to the prediction of the next word in a story.

[1] `http://www.mturk.com`, an online question answering service.

2.2.6 Adjective Noun Composition in the Brain

There has been some exploration of brain activation patterns in response to adjective processing by normal adults. Several studies utilizing MEG recordings have implicated right and left anterior temporal lobes (RATL and LATL) as well as ventro-medial prefrontal cortex (vmPFC). Adjective-noun pairs elicit increased neural activity when compared to word lists or non-words paired with nouns, with activity significantly higher in in LATL (184-255 ms), vmPFC (331-480 ms), and RATL (184 246 ms and 329 403 ms) (Bemis and Pylkkänen, 2011). When comparing a compositional picture naming task to a non-compositional picture naming task, Bemis and Pylkkänen (2013a) found differences in the magnitude of activation in LATL. Due to the timing of these effects, Bemis and Pylkkänen hypothesize that the activity in vmPFC is related to semantic processes, and that LATL activity could be due to the either the syntactic or semantic demands of composition.

Adjective-noun composition in fMRI was also explored by Chang et al. (2009b) with 12 adjective-noun pairs and semantic vectors comprised of verb co-occurrence statistics, as described in Section 2.1.1. They showed that, in terms of R^2 (regression coefficient of determination), a multiplicative model of composition outperformed an additive composition model, and also the adjective or the noun's semantic vector. However, in terms of ranking the predicted brain activation under the learned model by distance to the true brain activation, the additive, multiplicative and noun-only model were all within 2 percentage points of each other.

Baron and Osherson (2011) studied the semantic composition of adjective noun phrases using fMRI and a visual stimuli task. The stimuli was faces of young or old males (boys and men) and young or old females (girls and women). In the scanner, the faces were presented in blocks. For each block within the experiment, subjects were given a category (e.g. girl) and asked determine if each of the stimuli faces was a member of that category. Thus, for each block the face stimuli were the same, and only the concept being matched differed. Thus, any differences in activation can be attributed only to the matching task, and not to the stimuli. Baron and Osherson then created conceptual maps by learning regressors to predict brain activity based on the age (young or old) and gender of the matching task. Baron and Osherson found that the activation of a composed concept (e.g. young male) could be estimated by the multiplication or addition of adjective (e.g. young) and noun (e.g. boy) maps. They claim that multiplication "conforms better to the common usage of these concepts", presumably because it requires the activation of overlapping of brain areas rather than simple activation of disjoint areas. Areas of the brain that could be approximated well with an additive function were widespread and covered frontal, parietal, occipital and temporal lobes, whereas the multiplicative function was useful for predicting just to the left anterior temporal lobe (LATL).

Summary of Language in the Brain

How do the experimental results for semantic composition relate to the models of Semantic Unification previously discussed? If the syntactic form is held constant, as in adjective noun phrases or simple noun-verb-noun sentences, the combinatorial syntactic processes involved in language would be identical in the brain. However, when the semantics of the sentence changes due to different words or differing context, the semantic retrieval/memory and unification processes will

also change, resulting in differential brain activity.

In Bemis and Pylkkänen's work, the semantic content of the words is constant, but the task (combining words into phrases vs not), and thus the processing, differs. Their findings show increased activation in LATL, RATL and vmPFC, which implies that the combinatorial processes of adjective noun composition are at least partially handled in these areas. Hickok and Poeppel (2007) also hypothesize that the anterior temporal lobe is involved in composition, though they localize it to a slightly more medial temporal location. The vmPFC is not included in the Hickok model.

Recall the extremely distributed nature of semantics reported by Huth et al. (2012). Their finding implies that semantic portion of semantic composition will likely occur in many places in the brain, even if composition is mediated by areas of the temporal lobe. This is congruent with the additive model of Baron and Osherson (2011). Perhaps the temporal lobe acts like the conductor of an orchestra, and each semantic area is an instrument. Signals are sent by the conductor to raise or lower particular elements of the orchestra, or causes specific areas to begin to play in synchrony. This is a metaphor for the way the brain could encode changes in semantics due to composition, bringing the activation of brain areas up or down, or causing areas to work in synchrony to encode meaning altered by context.

2.3 Summary of Related Work

Semantic representations have been explored by both neuroscientists and computational linguists. Linguistic resources have helped neuroscientists map semantics onto the brain. So far, results on the semantic aspects of semantic composition are few. This thesis attempts to answer several questions to that end. For the study of language in the brain:

- In what form are semantic representations held in mind while reading new words?
- When is the output of semantic composition available neurally?
- Are compositional semantics encoded in the same areas of the brain responsible for the actual coordination of semantic composition (anterior temporal lobe, inferior frontal gyrus)?

For the study of composition using corpora:

- Can a corpus-based model of semantic composition be improved by incorporating the notion of composition into the model itself?
- How can the interpretability of a compositional model aid in our exploration of semantic composition?

For joining work from computational linguistics and neurolinguistics:

- Previous work has shown that brain- and corpus-based models of semantics are consistent (Mitchell et al., 2008; Murphy et al., 2012a). Are brain- and corpus-based representations of semantics also complementary? In other words, is there information available in brain data that can improve a purely corpus-based models of semantic, and vice versa?

The answers to these questions bring new insights into the semantic elements of semantic composition, and the role that brain and corpus data can play in the study of semantics.

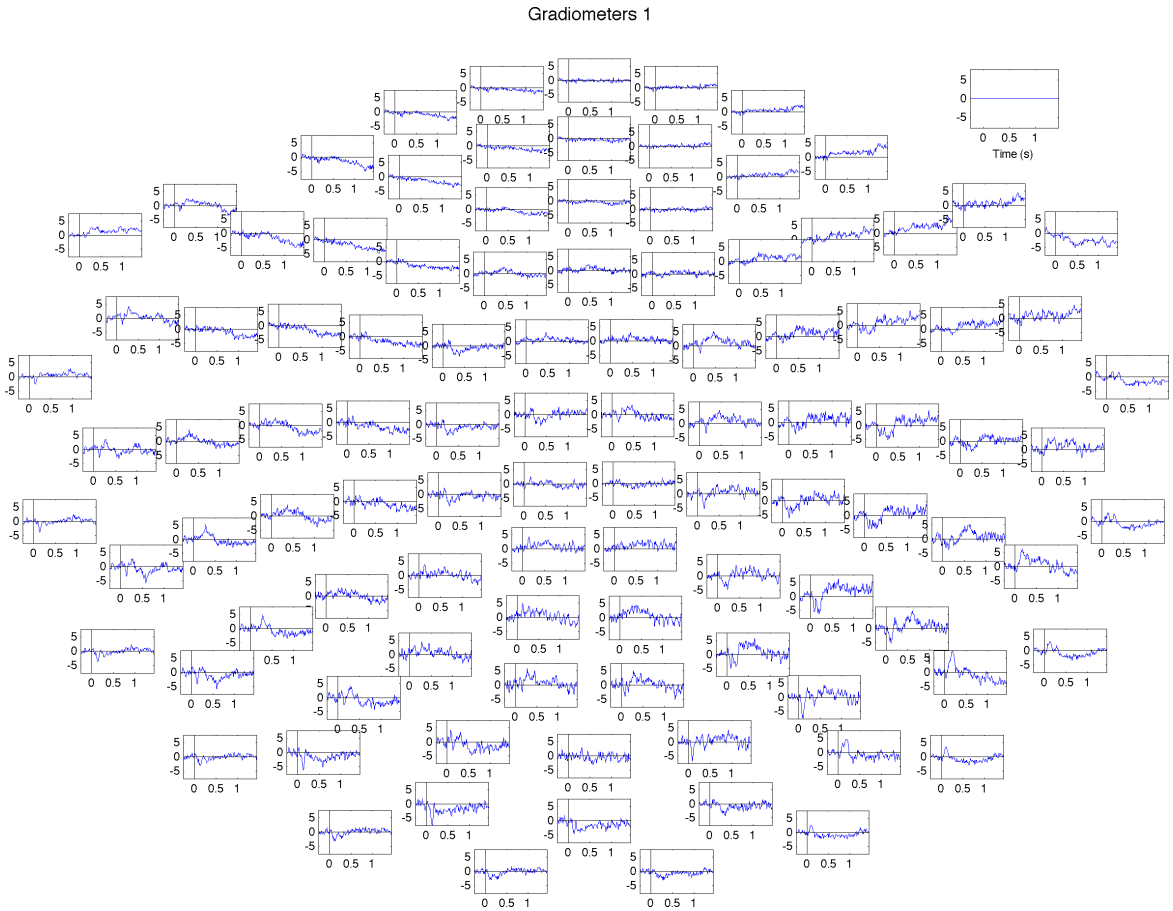


Figure 2.3: An example MEG recording averaged over 20 repetitions of a person reading the word “bear”. Each sub-plot represents the recording from the first gradiometer at one of the 102 sensor positions on the MEG helmet. For simplicity, the other 204 sensor recordings are not shown. In this diagram, the helmet is oriented as if we are looking down on it from above. The nose of the subject points to the top of the figure, and the left side of figure corresponds to the left hand side of the subject. Time is along the x axis of each plot and the y-axis corresponds to the gradient of the magnetic field in 10^{-3} T/cm.

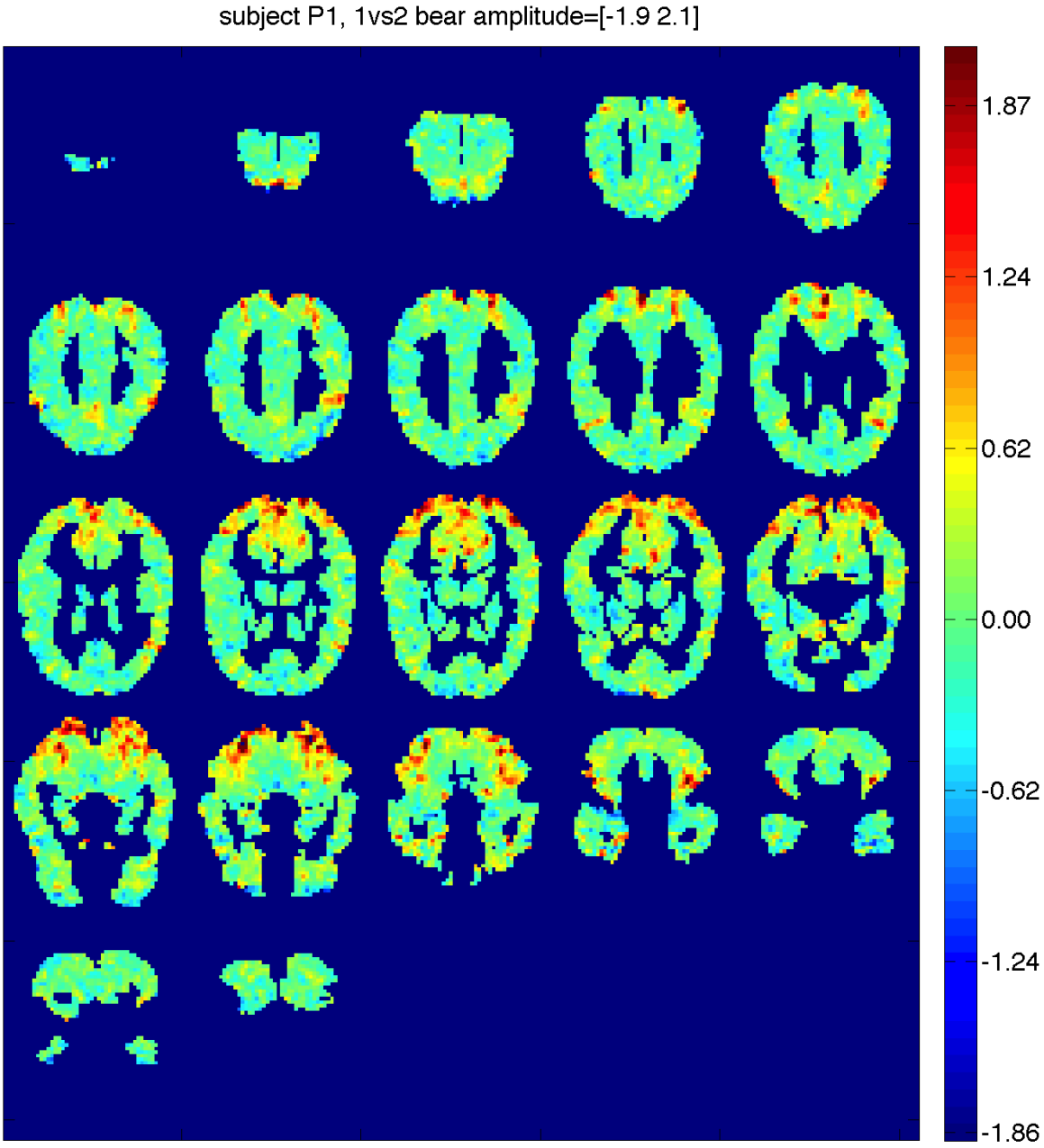


Figure 2.4: An fMRI image averaged over 6 repetition of a person reading the word "bear". An fMRI image is 3D, here shown in slices progressing from the top of the head (top left) to the bottom of the head (bottom right). In each slice, the front of the head points towards the bottom of the figure, and the right side of the subject is shown on the left side of the each image (as if we are viewing the brain of a subject laying face down, from the top of their head). The color of each voxel (pixel in brain space) represents the percent change over baseline of the BOLD response in that brain area.

Chapter 3

An Interpretable Model of Semantic Composition

This work is published as "A Compositional and Interpretable Semantic Space" (Fyshe et al., 2015).

Vector Space Models (VSMs) of Semantics are useful tools for exploring the semantics of single words. They can also be used to explore how the semantic composition of words leads to phrasal meaning. While many models can estimate the meaning (i.e. vector) of a phrase, few do so in an interpretable way. In this chapter, we introduce a method (CNNSE) that allows word and phrase vectors to adapt to the notion of composition. A VSM learned with our method is both interpretable and outperforms previously explored semantic composition methods. Interpretability is a powerful tool for exploring how words interact to create phrasal semantics; we leverage interpretability to analyze the performance of our model on several composition tasks.

3.1 Introduction

Vector Space Models (VSMs) are models of word semantics that are often built with word usage statistics derived from corpora. VSMs have been shown to closely match human perceptions of semantics for a variety of tasks (for an overview see Sahlgren (2006a), Chapter 5). Recently, VSMs have graduated beyond single words and have been used to study semantic composition: how words combine to create phrasal semantics (Mitchell and Lapata, 2010; Baroni and Zamparelli, 2010; Socher et al., 2012; Turney, 2012). Much of this work has focused on recreating phrasal vectors as functions of the vectors for the phrase's words, a task which can be performed fairly accurately.

The nature of composition has been explored along two main axes. Firstly, some work has explored different types of composition functions (Mitchell and Lapata, 2010) including higher order functions (such as matrices) (Baroni and Zamparelli, 2010). Secondly, some work has considered the types of corpus-derived information most useful for semantic composition (Turney, 2012; Fyshe et al., 2013). Still, many VSMs act like a black box - it is unclear what the

dimensions of a VSM represent (save for broad classes of corpus statistic types) and what the application of a composition function to those dimensions entails. Even popular methods that learn higher order functions for composition (Socher et al., 2012; Baroni and Zamparelli, 2010) lack interpretability; since individual dimensions cannot be assigned clear meaning, it is difficult to reason about how they might be combined via a higher-order function.

This chapter introduces a new method, Compositional Non-negative Sparse Embedding (CNNSE), that creates an interpretable VSM which takes semantic composition into account. In contrast to many other VSMs, our method learns a VSM that is tailored to suit the semantic composition function. The learned latent space (VSM) and composition function can be used to accurately predict the vectors of held out phrases. CNNSE is an extension of Non-Negative Sparse Embedding (NNSE) (Murphy et al., 2012b), an algorithm shown to produce VSMs in which the dimensions have clear and coherent meanings. Such interpretability allows for deeper exploration of semantic composition than previously possible.

This chapter begins with an overview of the CNNSE algorithm. We follow with empirical results that show our VSM produces:

1. better approximations of phrasal semantics when compared to semantic models that do not consider composition,
2. more interpretable dimensions than the typical VSM,
3. phrasal representations where the interpretable semantic dimensions more closely match the phrase meaning.
4. Composed representations that outperform previous methods on a phrase similarity dataset.

3.2 Method

Latent Semantic Analysis (LSA) (Deerwester et al., 1990) uses Singular Value Decomposition (SVD) to create a compact VSM. LSA models are incredibly useful and have been adopted for a variety of tasks including judging essay quality (Landauer and Laham, 1997) and estimating synonymy (Landauer and Dumais, 1997b). However, SVD often produces matrices where, for the vast majority of the dimensions, it is difficult to interpret what a high or low score entails for the semantics of a given word. In addition, the SVD factorization does not use of the phrasal relationships between the input words when compressing dimensions.

Our method extends Non-negative Sparse Embeddings (NNSEs) (Murphy et al., 2012b). NNSE increases interpretability by introducing sparsity and non-negativity constraints into a matrix factorization algorithm. The result is a VSM with extremely coherent dimensions, as quantified by a behavioral task (Murphy et al., 2012b). The output of NNSE is a matrix with rows corresponding to words and columns corresponding to latent dimensions.

To interpret a particular latent dimension, we can examine the words with the highest numerical values in that dimension (i.e. identify rows with the highest values for a particular column). For example, one dimension of an NNSE has top scoring words *sofa, couch, and mattress*, all of which are soft furniture items. We will refer to this word list as the dimension's *interpretable summarization*. To interpret the meaning of a word, we can examine the interpretable summarizations for the dimensions with the highest numerical value for that word (i.e. choose dimensions

with maximal score for a particular row). For example, in an NNSE VSM, the dimensions with the highest scores for the word *castle* have interpretable summarizations:

1. palace, castle, monastery
2. towers, tower, arches
3. towns, cities, town
4. dynasty, empire, emperors

conveying a castle's structural as well as royal facets. Table 3.7 has the interpretable summarization for the top scoring dimensions of a few example words and phrases.

NNSE is an algorithm which seeks a lower dimensional sparse representation for w words using the c-dimensional corpus statistics in a matrix $X \in \mathbb{R}^{w \times c}$. NNSE minimizes the following objective function:

$$\underset{A,D}{\operatorname{argmin}} \sum_{i=1}^{w} \left\| X_{i,:} - A_{i,:} \times D \right\|^2 + \lambda_1 \left\| A \right\|_1 \tag{3.1}$$

$$\text{st: } D_{i,:} D_{i,:}^T \leq 1, \forall\ 1 \leq i \leq \ell \tag{3.2}$$

$$A_{i,j} \geq 0,\ 1 \leq i \leq w,\ 1 \leq j \leq \ell \tag{3.3}$$

where $A_{i,j}$ indicates the entry at the ith row and jth column of matrix A, and $A_{i,:}$ indicates the ith row of the matrix. The solution includes a matrix $A \in \mathbb{R}^{w \times \ell}$ that is sparse, non-negative, and represents word semantics in an ℓ-dimensional latent space. $D \in \mathbb{R}^{\ell \times c}$ is the encoding of corpus statistics in the latent space. The L_1 constraint encourages sparsity in A; λ_1 is a hyperparameter. Equation 3.2 constrains D to eliminate solutions where the norm of A is made arbitrarily small by making the norm of D arbitrarily large. Equation 3.3 ensures that A is non-negative. Together, A and D factor the original corpus statistics matrix X in a way that minimizes reconstruction error while respecting sparsity and non-negativity constraints. One may tune ℓ and λ_1 to vary the sparsity of the final solution. Though we will not explore it here, ℓ can be greater than c but must be less than w.

Murphy et al. (2012b) solved this system of constraints using the Online Dictionary Learning algorithm described in Mairal et al. (2010). Though Equations 3.1-3.3 represent a non-convex system, when solving for A with D fixed (and vice versa) the loss function is convex. Mairal et al. break the problem into two alternating optimization steps (solving for A and$D$) and find the system converges to a stationary solution. The solution for A is found with a LARS implementation for lasso regression; D is found via gradient descent. Though the final solution may not be globally optimal, their method is capable of handling large amounts of data and has been shown to produce useful solutions in practice (Mairal et al., 2010; Murphy et al., 2012b).

We add an additional constraint to the NNSE loss function that allows us to learn a latent representation of semantics that respects the notion of semantic composition. As we will see, this change to the loss function has a huge effect on the learned latent space and its usefulness for studying semantic composition. Just as the L_1 regularizer can have a large impact on sparsity, our composition constraint represents a considerable change in composition compatibility.

Consider a phrase p made up of words i and j. In the most general setting, the following composition constraint could be applied to the rows of matrix A from Equation 3.1 corresponding

to p, i and j:

$$A_{(p,:)} = f(A_{(i,:)}, A_{(j,:)}) \tag{3.4}$$

where f is some composition function. The composition function constrains the space of learned latent representations $A \in \mathbb{R}^{w \times \ell}$ to be those solutions that are compatible with the composition function defined by f. Incorporating f into Equation 3.1 we have:

$$\begin{aligned} \underset{A,D,\Omega}{\operatorname{argmin}} \sum_{i=1}^{w} \left\|X_{i,:} - A_{i,:} \times D\right\|^2 + \lambda_1 \left\|A\right\|_1 + \\ \lambda_c \sum_{\substack{\text{phrase } p, \\ p=(i,j)}} \left(A_{(p,:)} - f(A_{(i,:)}, A_{(j,:)})\right)^2 \end{aligned} \tag{3.5}$$

Where each phrase p is comprised of words (i, j) and Ω represents all parameters of f that may need to be optimized. We have added a squared loss term for the composition function, and a new regularization parameter λ_c to weight the importance of respecting composition. We call this new formulation Compositional Non-Negative Sparse Embeddings (CNNSE). Some examples of the interpretable representations learned by CNNSE for adjectives, nouns appear in Table 3.7.

There are many choices for f: addition, multiplication, dilation, etc. (Mitchell and Lapata, 2010). Here we choose f to be weighted addition because it has has been shown to work well for adjective noun and noun noun composition (Mitchell and Lapata, 2010; Dinu et al., 2013), and because it leads to a formulation that lends itself well to optimization. Weighted addition is:

$$f(A_{(i,:)}, A_{(j,:)}) = \alpha A_{(i,:)} + \beta A_{(j,:)} \tag{3.6}$$

This choice of f requires that we simultaneously optimize for A, D, α and β. However, α and β are simply constant scaling factors for the vectors in A corresponding to adjectives and nouns. For adjective-noun composition, the optimization of α and β can be absorbed by the optimization of A. For models that include noun-noun composition, if α and β are assumed to be absorbed by the optimization of A, this is equivalent to setting $\alpha = \beta$.

We can further simplify the loss function by constructing a matrix B that imposes the composition by addition constraint. B is constructed so that for each phrase $p = (i, j)$:

$$\begin{aligned} B_{(p,p)} &= 1 \\ B_{(p,i)} &= -\alpha \\ B_{(p,j)} &= -\beta \end{aligned}$$

For our models, we use $\alpha = \beta = 0.5$, which serves to average the single word representations so that we avoid solutions where phrases have non-zero elements twice as large as single words. The matrix B allows us to reformulate the loss function:

$$\underset{A,D}{\operatorname{argmin}} \frac{1}{2}\left\|X - AD\right\|_F^2 + \lambda_1 \left\|A\right\|_1 + \frac{1}{2}\lambda_c \left\|BA\right\|_F^2 \tag{3.7}$$

B acts as a selector matrix, subtracting from the latent representation of the phrase the average latent representation of the phrase's constituent words. We have incorporated the factor $\frac{1}{2}$, which does not change the optimal solution, but simplifies the derivations in the following section.

We now have a loss function that is the sum of several convex functions of A: squared loss, L_1 regularization and the composition constraint. This sum of sub-functions is the format required for the alternating directions method of multipliers (ADMM) (Boyd, 2010). ADMM substitutes a dummy variable z for A in the sub-functions:

$$\underset{A,D}{\operatorname{argmin}} \frac{1}{2}\left\|X - AD\right\|_F^2 + \lambda_1\left\|z_1\right\|_1 + \frac{1}{2}\lambda_c\left\|Bz_c\right\|_F^2 \tag{3.8}$$

$$\text{st: } A = z_1 \tag{3.9}$$

$$A = z_c \tag{3.10}$$

$$D_{i,:}D_{i,:}^T \leq 1, \forall\ 1 \leq i \leq \ell \tag{3.11}$$

$$A_{i,j} \geq 0,\ 1 \leq i \leq w,\ 1 \leq j \leq \ell \tag{3.12}$$

Equations 3.9 and 3.10 ensure that the dummy variables match A; ADMM uses an augmented Lagrangian to incorporate and relax these new constraints. The augmented Lagrangian for the above optimization problem above is:

$$\begin{aligned} L_\rho(A, z_1, z_c, u_1, u_c) = \frac{1}{2}\left\|X - AD\right\|_F^2 + \lambda_1\left\|z_1\right\|_1 + \frac{1}{2}\lambda_c\left\|Bz_c\right\|_F^2 + \\ u_1(A - z_1) + u_c(A - z_c) + \frac{\rho}{2}(\left\|A - z_1\right\|_2^2 + \left\|A - z_c\right\|_2^2) \end{aligned} \tag{3.13}$$

We optimize for A, z_1 and z_c separately, and then update the dual variables (see Algorithm 2 for solutions and updates). ADMM has nice convergence properties for convex functions, as we have when solving for A. Code for ADMM is available online[1]. ADMM is used when solving for A in the Online Dictionary Learning algorithm, solving for D remains unchanged from the NNSE implementation (see Algorithm 1).

To begin, will explore CNNSE for adjective-noun composition. In Section 3.4 we will use CNNSE to model adjective-noun and noun-noun composition simultaneously.

We use the weighted addition composition function because it performed well in previous work (Mitchell and Lapata, 2010; Dinu et al., 2013), maintains the convexity of the loss function, and is easy to optimize. In contrast, element-wise multiplication or dilation composition functions leads to a much more complex non-convex optimization problem which cannot be solved using ADMM. Though not explored here, we hypothesize that A could be molded to respect many different composition functions. However, if the chosen composition function does not maintain convexity, finding a suitable A may prove challenging. We also hypothesize that even if the chosen composition function is not the "true" composition function (whatever that may be), the fact that A can change to suit the composition function may compensate for this mismatch. This has the flavor of variational inference for bayesian methods: an approximation in place of an intractable problem often yields better results with limited data, in less time.

[1] `http://www.stanford.edu/~boyd/papers/admm/`

Algorithm 1 CNNSE

Input: $X, B, \lambda_1, \lambda_c$
Randomly initialize A, D
prevL $\Leftarrow 0$
curL $\Leftarrow \frac{1}{2}\left\|X - AD\right\|_F^2 + \lambda_1\left\|A\right\|_1 + \frac{1}{2}\lambda_c\left\|BA\right\|_F^2$
while (prevL - currL) $\leq$ prevL$*10^{-3}$ **do**
  $A \Leftarrow$ ADMM$(D, X, B, \lambda_1, \lambda_c)$
  $D \Leftarrow$ gradientDescent(D, X, A)
  prevL $\Leftarrow$ curL
  curL $\Leftarrow \frac{1}{2}\left\|X - AD\right\|_F^2 + \lambda_1\left\|A\right\|_1 + \frac{1}{2}\lambda_c\left\|BA\right\|_F^2$
end while
return A, D

3.3 Data and Experiments

We use the semantic vectors made available by Fyshe et al. (2013)[2]. We used the 1000 dependency SVD dimensions, which were shown to perform well for composition tasks. The dependency features were compiled from a 16 billion word subset of ClueWeb09 (Callan and Hoy, 2009). Features are tuples consisting of two part-of-speech tagged words and the dependency relationship between them; the feature value is the pointwise positive mutual information (PPMI) for the tuple:

$$\text{PPMI} = \max\left(log\left(\frac{p(x,y)}{p(x)p(y)}\right), 0\right) \tag{3.14}$$

where x is a word, y is a word-dependency pair, and p is the probability as estimated over the corpus. The Fyshe et al. (2013) dataset is comprised of 54,454 words and phrases. We randomly split the approximately 14,000 adjective noun phrases into a train (2/3) and test (1/3) set. From the test set we removed 200 randomly selected phrases as a development set for parameter tuning.

NNSE has one parameter to tune (λ_1); CNNSE has two: λ_1 and λ_c. In general, these methods are not overly sensitive to parameter tuning, and searching over orders of magnitude will suffice. We found the optimal settings for NNSE were $\lambda_1 = 0.05$, and for CNNSE $\lambda_1 = 0.05, \lambda_c = 0.5$. Too large a value for λ_1 lead to overly sparse solutions, too small negatively impacted interpretability. We set $\ell = 1000$ for both NNSE and CNNSE and altered sparsity by tuning only λ_1 (lower ℓ essentially guarantees sparsity by ensuring that omitted dimensions $> \ell$ are all 0).

3.3.1 Phrase Estimation

We have incorporated the notion of composition into our loss function, and so the learned latent space could be a more accurate predictor of compositional semantics. To test the ability of each model to estimate phrase semantics we trained models on the training set, and then used the learned model to predict the withheld phrase rows of X. We sort the vectors for the test phrases,

[2] http://www.cs.cmu.edu/~afyshe/papers/conll2013/

Algorithm 2 ADMM solution for augmented Lagrangian in equation 3.13

Input: $D, X, B, \lambda_1, \lambda_c$
{Lagrangian parameter}
$\rho \Leftarrow 1$
{Dummy Variables}
$z_1 \Leftarrow 0_{w,\ell}$
$z_c \Leftarrow 0_{w,\ell}$
{Dual Variables}
$u_1 \Leftarrow 0_{w,\ell}$
$u_c \Leftarrow 0_{w,\ell}$
$dti \Leftarrow DD^T + 2 * \rho * I_m$
while not converged **do**
$A \Leftarrow \left(XD^T + \rho(z_1 + z_c) - (u_1 + u_c)\right)/dti$
$z_c \Leftarrow (\rho * A + u_c)/(\lambda_c * (B' * B) + \rho * I_w)$
$\gamma \Leftarrow A + u_1/\rho$
$\kappa \Leftarrow \lambda_1/\rho$
{Soft Threshold Operator for L_1 constraint} {$(a)_+$ is shorthand for max(0,a)}
$z_1 = (\gamma - \kappa)_+ - (-\gamma - \kappa)_+$
{Update Dual Variables}
$u_1 = u_1 + \rho * (A - z_1)$
$u_c = u_c + \rho * (A - z_c)$
end while
return A

X_{test}, by their cosine distance to the predicted phrase $\hat{X}_{(p,:)}$. We report two measures of accuracy. The first is median rank accuracy, where rank accuracy is:

$$\text{rank accuracy} = 100 \times \left(1 - \frac{r}{P}\right) \tag{3.15}$$

r is the position of the correct phrase in the sorted list of test phrases, and $P = |X_{test}|$ (the number of test phrases). The second measure is mean reciprocal rank (MRR), which proved to be a much more discriminative measure of phrasal prediction. MRR is:

$$\text{MRR} = 100 \times \left(\frac{1}{P}\sum_{i=1}^{P}(\frac{1}{r})\right) \tag{3.16}$$

For both measures a perfect score is 100. If the correct phrase is always ranked second median rank accuracy would be $100 \times (1 - (P-1)/P)) = 99.95$ for our test set. In contrast, MRR would be 50. If the correct phrase is always ranked 50th, median rank accuracy would be 98.85 and MRR would be 2. Thus, MRR places much more emphasis on ranking items close to the top of the list, and less on differences in ranking lower in the list.

We will compare to two other previously studied composition methods: weighted addition (**w. add**), and **lexfunc** (Baroni and Zamparelli, 2010). Weighted addition finds α, β to optimize

$$(X_{(p,:)} - \alpha X_{(i,:)} - \beta X_{(j,:)})^2 \tag{3.17}$$

Table 3.1: Median rank, mean reciprocal rank (MRR) and percentage of test phrases ranked perfectly (i.e. at the top of the sorted list of ~ 4600 test phrases) for four methods of estimating the corpus statistics X for phrases in the test set. w. add is weighted addition of SVD vectors, w. NNSE is weighted addition of NNSE vectors.

Model	Med. Rank	MRR	Perfect rank
w. add	99.89	35.26	20%
Lexfunc	99.65	28.96	20%
w. NNSE	99.80	28.17	16%
CNNSE	**99.91**	**40.64**	**26%**

Note that this optimization is performed over the training vectors of the SVD input X, rather than on the learned latent space A. To estimate X for a new phrase $p = (i, j)$ we compute

$$\hat{X}_{(p,:)} = \alpha X_{(i,:)} + \beta X_{(j,:)} \tag{3.18}$$

Lexfunc finds an adjective-specific matrix M_i that solves

$$X_{(p,:)} = M_i X_{(j,:)} \tag{3.19}$$

for all phrases $p =< i, j >$ for a given adjective i. We solved each adjective-specific problem with Matlab's partial least squares implementation, which uses the SIMPLS algorithm (Dejong, 1993). To estimate X for a new phrase $p = (i, j)$ we compute

$$\hat{X}_{(p,:)} = M_i X_{(j,:)} \tag{3.20}$$

We also optimized the weighted addition composition function over NNSE vectors, which we call **w. NNSE**. We compose the latent phrase vectors to estimate the held out $A_{(p,:)}$:

$$\hat{A}_{(p,:)} = \alpha A_{(i,:)} + \beta A_{(j,:)} \tag{3.21}$$

after optimizing α and β using the training set. For CNNSE, as in the loss function, $\alpha = \beta = 0.5$ so that the average of the word vectors approximates the phrase.

$$\hat{A}_{(p,:)} = 0.5 \times (A_{(i,:)} + A_{(j,:)}) \tag{3.22}$$

Crucially, w. NNSE estimates α, β *after* learning the latent space A, whereas CNNSE *simultaneously* learns the latent space A, while taking the composition function into account. Once we have an estimate $\hat{A}_{(p,:)}$ we can use the NNSE and CNNSE solutions for D to estimate the corpus statistics X.

$$\hat{X}_{(p,:)} = \hat{A}_{(p,:)} D \tag{3.23}$$

For all four methods, we sort the rows of matrix X_{test} by their cosine distance to $\hat{X}_{(p,:)}$ and calculate MRR and median rank accuracy of the correct phrase in the sorted list.

Results for the four methods appear in Table 3.1. Median rank accuracies were all within half a percentage point of each other. However, MRR shows a striking difference in performance. CNNSE has MRR of 40.64, more than 5 points higher than the second highest MRR score belonging to w. add (35.26). CNNSE ranks the correct phrase in the first position for 26% of phrases, compared to 20% for w. add. Lexfunc ranks the correct phrase first for 20% of the test phrases, w. NNSE 16%. So, while all models perform quite well in terms of rank accuracy, CNNSE is the clear winner when we consider MRR.

We were surprised to find that lexfunc performed relatively poorly in our experiments. Dinu et al. (2013) used simple unregularized regression to estimate M. We also replicated that formulation, and found phrase ranking to be worse when compared to the Partial Least Squares method described in Baroni and Zamparelli (2010). In addition, Baroni and Zamparelli use 300 SVD dimensions to estimate M. We found using 1000 SVD dimensions performed slightly better. We hypothesize that our difference in performance could be due to the difference in input corpus statistics (in particular the thresholding of infrequent words and phrases), or due to the fact that we did not specifically create the training and tests sets to evenly distribute the phrases for each adjective. If an adjective i appears only in phrases in the test set, lexfunc cannot estimate M_i using training data (a hindrance not present for other methods, which require only that the adjective appear in the training data). To compensate for this possibly unfair train/test split, the results in Table 3.1 are calculated over only those adjectives which could be estimated using the training set. Though the results reported here are not as high as previously reported, lexfunc was found to be only slightly better than w. add for adjective noun composition (Dinu et al., 2013). CNNSE outperforms w. add by a large margin, so even if Lexfunc could be tuned to perform at previous levels on this dataset, CNNSE would likely still dominate.

None of the models explored here are perfect. Even the top scoring model, CNNSE, only identifies the correct phrase for 26% of the test phrases. When a model makes a "mistake", it is possible that the top-ranked word is a synonym of, or closely related to, the actual phrase. To evaluate mistakes, we chose test phrases for which all 4 models are incorrect and all 4 rank a different phrase as their top choice. We believe these examples likely represent the most difficult phrases to estimate. We then asked Mechanical Turk[3] users to evaluate the mistakes. We presented the 4 mistakenly top-ranked phrases to Mechanical Turk users, who were asked to choose the one phrase most related to the actual test phrase.

We randomly selected 200 such phrases and asked 5 Mechanical Turk users to evaluate each, paying \$0.01 per answer. We report here the results for questions where a majority (3) of users chose the same answer (82% of questions).

Table 3.2 shows the Mechanical Turk evaluation of model mistakes. CNNSE and lexfunc make the most reasonable mistakes, having their top-ranked phrase chosen as the most related phrase 35.4% and 31.7% of the time, respectively. This makes us slightly more comfortable with our phrase estimation results (Table 3.1); though lexfunc is does not reliably predict the correct phrase, it often chooses a close approximation. The mistakes from CNNSE are chosen slightly more often than lexfunc, indicating that CNNSE also has the ability to reliably predict the correct phrase, or a related phrase.

[3] `http://mturk.com`

Table 3.2: Results for a Mechanical Turk experiment to determine the model that makes the most reasonable mistakes in phrase ranking. To evaluate mistakes, we chose phrases for which all 4 models are incorrect and all 4 rank a different phrase as their top choice. Mechanical Turk users were asked to choose the mistake phrase that was the closest match to the target phrase.

Model	Predicted phrase deemed closest match to actual phrase
w. add	21.3%
Lexfunc	31.7%
w. NNSE	11.6%
CNNSE	35.4%

3.3.2 Interpretability

Ranking lists of phrases is a common benchmark for testing models of semantic composition. As our results have shown, many systems perform very well at this task. Though our improvement in MRR is compelling, we seek to explore the meaning encoded in the word space features. We turn now to the *interpretation* of phrasal semantics and semantic composition.

One of the advantages of NNSE is that it produces interpretable semantic representations. For each dimension of the learned latent representation (columns of A) we can select the highest scoring words to create interpretable summarization for that dimension. Due to the sparsity and non-negativity constraints, NNSE produces dimensions with very coherent definitions (Murphy et al., 2012b). Murphy et al. used an intruder task to quantify the interpretability of semantic dimensions. The intruder task presents a human user with a list of words, and they are to choose the one word that does not belong in the list. For example, from the list

- red
- green
- desk
- pink
- purple
- blue

it is clear to see that the word "desk" does not belong in the list of colors. To create questions for the intruder task, we selected the top 5 scoring words in a particular dimension, as well as a low scoring word from that same dimension such that the low scoring word is also in the top 10th percentile of some other dimension. Like the word "desk" in the example above, this low scoring word is called the *intruder*, and the human subject's task is to select the intruder from a shuffled list of the 6 words. Mechanical Turk was used to collect responses for this task. Five Mechanical Turk users answered each question, each paid $0.01 per answer. A high percentage of intruders detected by Mechanical Turk users indicates that the latent semantic representation groups semantically similar words in a human-interpretable fashion. We chose 100 questions for each of the NNSE, CNNSE and SVD representations. Lexfunc was not evaluated for interpretability because it is unclear how the matrix M could be easily interpreted. However, the

Table 3.3: Results from Mechanical Turk task to evaluate the interpretability of the learned semantic dimensions. Intruders detected is the percentage of questions for which the majority response was correct, and human agreement is the percentage of questions for which a majority of the 5 Mechanical Turk users chose the same response from the list of 6 words. As in (Murphy et al., 2012b), SVD dimensions are highly uninterpretable, NNSE and CNNSSE dimensions are represent highly consistent concepts.

Method	Intruders Detected	Human Agreement
SVD	17.6%	74%
NNSE	86.2%	94%
CNNSE	88.9%	90%

output of lexfunc is the SVD representation X, so the SVD interpretation could be considered a proxy for lexfunc interpretability.

Results for our interpretability experiment are presented in Table 3.3. Consistent with previous studies, NNSE provides a much more interpretable latent representation than SVD. We find that the additional composition constraint used in CNNSE has maintained the interpretability of the learned latent space. Because intruders detected is higher for CNNSE, but agreement amongst Mechanical Turk users is higher for NNSE, we consider the interpretability results for the two methods essentially equivalent. Note that SVD interpretability is close to chance ($1/6 = 0.1667$).

The dimensions of NNSE and CNNSE are comparably interpretable. But, has the composition constraint in CNNSE resulted in better phrasal representations? To test this, we randomly selected 200 phrases, and then identified the top scoring dimension for each phrase in both the NNSE and CNNSE models. We presented Mechanical Turk users with the interpretable summarizations for a phrase as learned by CNNSE and NNSE. Users were asked to select the list of words that was most closely related to the target phrase. Mechanical Turk users could also select that neither list was related, or that the lists were equally related to the target word. We paid \$0.01 per answer and had 5 users answer each question. We report results for phrases where at least 3 of 5 users selected the same answer (78% questions). Results for this task appear in Table 3.4. CNNSE phrasal representations are found to be much more consistent, receiving a positive evaluation almost twice as often as NNSE.

Together, these results show that CNNSE representations maintain the interpretability of NNSE representations, while improving on the ability of the VSM to model semantic composition. Table 3.7 shows a few examples of adjective, noun and actual/estimated phrasal representations.

3.3.3 Evaluation on Behavioral Data

We now compare the performance of various composition methods on the adjective-noun phrase similarity dataset from Mitchell and Lapata (2010). This dataset is comprised of 108 adjective-noun phrase pairs split into high, medium and low similarity groups. An example of a high similarity phrase pair is *general principle, basic rule*, a low similarity phrase pair: *large quantity, good place*. Similarity scores from 18 human subjects were collected for each phrase pair. We

Table 3.4: Results from an experiment to compare the interpretable phrasal representations of CNNSE and NNSE. Mechanical Turk users were shown the interpretable summarization for the top scoring dimension for target phrases. Representations from CNNSE and NNSE were shown side by side and Mechanical Turk users were asked to choose the list most related to the phrase. Users could also select that neither list was a good match to the phrase, or that the lists were equally good.

Model	Model representation deemed most consistent with phrase
CNNSE	54.5%
NNSE	29.5%
Both	4.5%
Neither	11.5%

average together the 18 scores to create one similarity score per phrase pair. We then compute the cosine similarity between the composed phrasal representations of each phrase pair under each compositional model. We report the correlation of the cosine similarity measures to the behavioral scores. We withheld 12 of the 108 questions for parameter tuning, four randomly selected from each of the high, medium and low similarity groups.

Table 3.5 shows the correlation of each model's similarity scores to behavioral similarity scores. Again, Lexfunc performs poorly. This is probably attributable to the fact that there are, on average, only 39 phrases available for training each adjective in the dataset, whereas the original Lexfunc study had at least 50 per adjective (Baroni and Zamparelli, 2010). CNNSE is the top performer, followed closely by weighted addition. Interestingly, weighted NNSE correlation is lower than CNNSE by nearly 0.15, which shows the value of allowing the learned latent space to conform to the desired composition function.

Though the difference on this task between weighted addition and CNNSE is small, CNNSE has the additional advantage of interpretability. To illustrate this, we have created a web page to explore this dataset under the CNNSE model. The page `http://www.cs.cmu.edu/~fmri/papers/naacl2015/cnnse_mitchell_lapata_all.html` displays the phrase pairs from Mitchell and Lapata (2010) sorted by average similarity score (as judged by the human subjects). For each phrase in the pair we show a summary of the CNNSE composed phrase meaning, as in Section 3.2. The scores of the 10 top scoring dimensions are displayed in descending order. Each dimension is described by its interpretable summarization. Figure 3.1 shows an example phrase pair from the web page. As one scrolls down the page, the similarity scores increase, and the number of dimensions shared between the phrase pairs (highlighted in red) increases. Some phrase pairs with high similarity scores share no top scoring dimensions. Because we can interpret the dimensions, we can begin to understand how the CNNSE model is failing, and how it might be improved.

For example, the phrase pair judged most similar by the human subjects, but that shares none of the top 10 dimensions in common, is "large number" and "great majority" (behavioral similarity score 5.61). Upon exploration of the phrasal representations, we see that the representation for "great majority" suffers from the multiple word senses of majority. Majority is often used

in political settings to describe the party or group with larger membership. We see that the top scoring dimension for "great majority" has top scoring words "candidacy, candidate, caucus", a politically-themed dimension. Though the representation is not incorrect for the word, one might argue that the themes in common between the two test phrases are not political. On the other hand, the second highest scoring dimension for "large number" is "First name, address, complete address". Here we see another case of the collision of multiple word senses, as this dimension is probably related to the phone numbers or house numbers, rather than the quantity-related sense of number. While it is satisfying that the word senses for majority and number have been separated out into different dimensions for each word, it is clear that both the composition and similarity functions used for this task are not gracefully handling the multiple word senses. To solve this problem, we could amend the similarity function to account for the fact that different dimensions represent different word senses. One could imagine partitioning the learned latent dimensions of A into sense-related groups and using the maximally correlated subset of the dimensions to score phrase pairs. It is the interpretability of CNNSE that allows us to do these analyses and to make these observations.

Table 3.5: Correlation of behavioral data to pairwise distances of vectors from several adjective-noun composition models. Behavioral data is from Mitchell and Lapata (2010).

Model	Correlation to behavioral data
w. add	0.5377
Lexfunc	0.1347
w. NNSE	0.4469
CNNSE	0.5923

3.4 Adjective-noun and noun-noun composition

Up until this point, the dataset used in this chapter contained only adjective noun phrases, but noun-noun phrases have similar properties, and their compositional nature is similar (Mitchell and Lapata, 2010; Turney, 2012). In this section, we train CNNSE on a dataset that contains both adjective-noun and noun-noun phrases courtesy of Turney (2012). We will compare Turney's *dual-space* composition system to our learned latent compositional representation. Can our system create an interpretable model that works well for both adjective-noun and noun-noun phrases?

Turney's dataset is based on a different corpus, and the statistics collected using the corpus are also different. This corpus is based on web pages that total 5×10^{10} words. The words for which statistics were collected are taken from the WordNet lexicon (Fellbaum, 1998). Turney explores two variants of corpus statistics for what he terms Domain and Function spaces. Domain space is based on the first noun found to the left and to the right of the target word. Function space is based on the first verb to the left and to the right of the target word. Both spaces use a context window of 7 words on either side of the word of interest, and if no noun or verb is found within

Figure 3.1: A example adjective-noun phrase similarity question from Mitchell and Lapata (2010). The header shows the two phrases of the example phrase pair. Below are the CNNSE representations for each composed phrase. Below are the scores for the top scoring dimensions of the composed phrase vector, as well as the interpretable summarizations each dimension. Results for all questions from the dataset can be viewed at `http://www.cs.cmu.edu/~fmri/papers/naacl2015/cnnse_mitchell_lapata_all.html`

early/JJ evening/NN **(behav sim score 2.78)**		previous/JJ day/NN	
0.3480	afternoon/NN afternoons/NNS evening/NN	0.1871	biennial/JJ future/JJ national/JJ
0.1479	fall/NN monday/NN next/JJ_Friday/NNP	0.1719	afternoon/NN afternoons/NNS evening/NN
0.0614	eleventh/JJ fifteenth/JJ fifth/JJ	0.0978	conflicting/JJ conflicting/NN contradictory/JJ
0.0532	all-day/JJ all-night/JJ celebratory/NN	0.0975	biweekly/JJ daily/JJ frequent/JJ
0.0489	Last/JJ_fall/NN Last/JJ_month/NN Last/JJ_week/NN	0.0934	eleventh/JJ fifteenth/JJ fifth/JJ
0.0395	banquet/NN conference/NN exhibition/NN	0.0655	conquered/JJ developing/JJ feudal/JJ
0.0288	ardent/JJ die-hard/JJ erstwhile/NN	0.0608	conditional/JJ definitive/JJ implied/JJ
0.0267	conquered/JJ developing/JJ feudal/JJ	0.0587	all-out/JJ decades-long/JJ ensuing/VBG
0.0215	anniversary/NN birthday/NN birthdays/NNS	0.0553	anniversary/NN birthday/NN birthdays/NNS
0.0202	getaway/NN getaway/NNP interlude/NN	0.0516	alleged/JJ attempted/JJ blatant/JJ

that window, no count is generated for the target word. Domain and Function space are meant to represent the topic-related and role-related aspects of a word. PPMI (pointwise positive mutual information) is used to normalize the counts. Then, SVD is applied to the matrix to compute the left-singular vectors and eigenvalues of the PPMI matrix.[4]

Turney develops a specialized similarity function to compare adjective-noun and noun-noun phrases to single words. The similarity function considers the domain and function statistics separately and combines them into one score. For phrase ab comprised of words a and b and some other word c Turney uses:

$$sim_1(ab, c) = geo(sim_d(a, c), sim_d(b, c), sim_f(b, c)) \tag{3.24}$$

$$sim_c(ab, c) = \begin{cases} sim_1(ab, c) & \text{if } a \neq c \text{ and } b \neq c \\ 0 & \text{otherwise} \end{cases} \tag{3.25}$$

Here, sim_d and sim_f are cosine similarity using the domain and function statistics for the given words, and geo is the geometric mean. To evaluate, Turney created a new task: multiple-choice noun-modifier composition questions. The query for each question is either an adjective-noun or noun-noun phrase, and the 7 possible answers are single words. The correct answer is a one word synonym for the noun phrase. The questions are quite difficult, as the possible answers contain both constituent words of the phrase, as well as synonyms for the constituent words and two randomly chosen nouns. For example, for a query phrase like "dog house" the correct answer

[4]Thank-you to Peter Turney for supplying the domain and function word vectors used to conduct these experiments, which allowed us to perform a fair comparison of methods.

is "kennel", and the other possible answers are "dog", "house", "canine", "dwelling", "effect", and "largeness". The dataset contains 2180 questions, with 680 questions used for parameter tuning. Turney tunes several parameters for his model: the number of domain and function dimensions used, and the power to which the eigenvalues are raised before multiplying with the left-singular vectors. We use the development set to choose λ_1 and λ_c for CNNSE.

Note that in Equation 3.25, Turney has hard-coded the fact that a phrase ab and word c should have zero similarity if the phrase contains word c. This side-steps a common pitfall for many adjective noun composition methods: the composed vector is often still very similar to one of its constituent words. Recall that each of the questions in this evaluation set contains both the adjective and the noun as options. Thus, without this special handling of constituent words, performance would suffer a great deal. For a fair and accurate comparison, we use sim_c as defined above, but re-define sim_1 to be the simple cosine distance between the composed phrase vector for ab and the single word c.

Table 3.6 shows the results on the multiple-choice noun-modifier composition questions using the domain and function vector space as input to both CNNSE and the specially designed dual-space similarity function. Results are shown both with and without the constituent word constraint of Equation 3.25. With the constraint, the dual-space model outperforms for both percentage of questions correct and MRR (as calculated over the 7 possible word answers). However, the performance is close and, as we will see, CNNSE allows us to interpret the results in a way that isn't possible in the original dual-space model. Without the constituent word constraint, CNNSE and dual-space have more similar performance (a difference of 16/1500 questions correct separates the two methods). This shows how important the constituent word constraint is to both methods. It should be noted that the dual-space model does not actually estimate a phrase vector, it simply sorts candidate words c by their sim_1 score to the words ab of a particular phrase (Equation 3.25). Our results show that learning a latent space that can adapt to a particular composition function can be nearly as fruitful as handcrafting a comparison function.

Turney has since improved on the results from the dual-space model with Super-Sim, an SVM trained to predict if a phrase and candidate word are synonymous. The input to the SVM are the domain and function vectors (from the dataset used in this section) for the words of the phrase, plus the logarithmic frequency of the words and the pointwise-positive mutual information between the two words. This method's performance is greatly improved over the dual-space model, answering 75.9% of questions correctly. In addition, the supervised nature of the task allows Turney to drop the constituent word constraint, as it is learned automatically. However, this method adds yet another level of complication to the model, making the exploration of composition even more difficult, as the comparison function is now a learned set of weights over largely uninterpretable vectors.

We can use CNNSE to explore the results from the multiple-choice noun-modifier composition questions. The web page at `http://www.cs.cmu.edu/~fmri/papers/naacl2015/cnnse_turney.html` shows 100 randomly chosen questions and their CNNSE representations. As in Section 3.3.3, CNNSE representations are summarized by the top scoring dimensions for a particular word (with scores), and each dimension is described by its top scoring words. Figure 3.2 shows an example question from the web page. On the left, in blue, are the noun phrases with their composed representations directly below. To the right of each noun phrase are the candidate answers, sorted by cosine similarity to the composed representation of the phrase. The

cosine similarity is shown in parenthesis next to each candidate answer. The correct answer is highlighted in green. The CNNSE representation of each candidate answer is also shown. For brevity, the adjective and noun representations are omitted, as they are essentially eliminated from the ranking by the constituent word constraint.

A quick scan through the CNNSE representations makes it clear that this task is very hard. The synonyms of the constituent words are strong distractors, and many of the correct answers are uncommon words or scientific terms. For example, some phrases and their correct answers: "wood coal": "lignite", "enlarged heart": "cardiomegaly", "peace lily": "spathiphyllum". One of the most common mistakes is to rank a synonym for a constituent word higher than the true answer, as can be seen in the example in Figure 3.2. This is a case where a supervised approach could be very useful, as an algorithm could learn that if a candidate word is too similar to one of the words of the phrase, but not the other, the candidate is not the correct answer. Still, this tuning of a comparison function side steps the actual issue of composition entirely.

3.5 Conclusion

In this chapter, we explored a new method to create VSMs that respects the notion of semantic composition. We found that our technique for incorporating phrasal relationship constraints produced a VSM that is more consistent with observed phrasal representations and with behavioral data.

Because this work extends a model that produces interpretable semantic representations, we were able to use our model to explore semantic composition in the context of our data. We found that, compared to an interpretable model that does not incorporate semantic composition, human evaluators judged the phrasal representations from our compositional model to be a better match to phrase meaning. We leveraged this improved interpretability to explore composition in the context of two previously published compositional tasks. We note that the collision of word senses often hinders performance on the behavioral data from Mitchell and Lapata (2010).

More generally, this chapter illustrates that incorporating constraints to represent the task of interest can improve a model's performance on that task. Additionally, incorporating such constraints into an interpretable model allows for a deeper exploration of performance in the context of evaluation tasks.

Table 3.6: Results for multiple-choice noun-modifier composition questions from Turney (2012). Percentage correct is the number of questions for which the correct answer was ranked in the top position. MRR is mean reciprocal rank for the rank-order of the answers.

	With constraint in Eq 3.25		Without constraint in Eq 3.25	
Model	Percentage correct	MRR	Percentage correct	MRR
dual-space	58.27	77.0	13.7	42.9
CNNSE	52.60	73.1	12.60	36.1

Figure 3.2: A example multiple-choice noun-modifier composition question from Turney (2012), containing the CNNSE representations for the composed phrase and the single word answers. The composed phrase (high spot) appears at the left, highlighted in blue. The correct answer (highlight) appears with a green background. Below each word/phrase are the scores for the top scoring dimensions for the word/phrase. Then, for each selected dimension the interpretable summarization is given, which reports the top scoring words in that dimension. A web page with 100 randomly selected questions is available at `http://www.cs.cmu.edu/~fmri/papers/naacl2015/cnnse_turney.html`

high spot		degree (0.0094)		place (0.0058)		highlight (0.0000)		immunosuppression (0.0000)		knacker (0.0000)	
0.0292	blight, cercospora, cercosporella,	0.3452	baccalaureate, bmus, doctoral,	0.0323	catalyst, centerpiece, cornerstone,	0.1616	analyze, assess, conceptualize,	0.1991	alpha-interferon, chemotherapy, chlorambucil,	0.2659	hakenkreuz, pfannkuchen, plastinate,
0.0212	championship, elimination tournament, playoff,	0.0310	arrest warrant, certificate, injunction,	0.0136	clomid, dilantin, glucophage,	0.1593	cascading menu, chooser, dialog,	0.1477	allogeneic, allograft, autograft,	0.1498	baum, frisch, fuchs,
0.0132	day school, kindergarten, prep school,	0.0164	billionth, centimeter, centimetre,	0.0056	disinclined, duty-bound, intend,	0.1544	co-ordinate, decentralise, finalise,	0.1086	acromegaly, hypercalcemia, hypernatremia,	0.1089	actinoid, adscititious, alveolate,
0.0089	decrease, decreased, decreasing,	0.0120	acceptability, adequacy, appropriateness,	0.0045	baccarat, betting, bettor,	0.1022	arouse, captivate, enlighten,	0.0952	arthropathy, dermatomyositis, glomerulonephritis,	0.0598	andres martinez, crapulent, dahna,
0.0065	alloy steel, austenitic, austenitic steel,	0.0119	alacrity, clearness, directness,	0.0042	beach house, beachfront, chalet,	0.0776	embody, encompass, exemplify,	0.0852	cytolytic, eosinophil, leukocyte,	0.0328	abfarad, abvolt, admass,

Table 3.7: A qualitative evaluation of CNNSE interpretable dimensions for several phrases and their constituent words. For each word or phrase the top 5 scoring dimensions are selected. Then, for each selected dimension the interpretable summarization is given, which reports the top scoring words in that dimension.

Adjective	Noun	Phrase	Estimated Phrase
negative aspects			
negative	aspects	negative aspects (observed)	negative aspects (estimated)
intruders, intrusions, overflows	facets, topics, different aspects	consequences, environmental consequences, serious consequences	facets, topics, different aspects
consequences, environmental consequences, serious consequences	underpinnings, arousal, implications	features, oddities, standard features	underpinnings, arousal, implications
instinctive, conditioned, oscillatory	features, oddities, standard features	intruders, intrusions, overflows	intruders, intrusions, overflows
indecent, unlawful, obscene	workings, truths, essence	facets, topics, different aspects	consequences, environmental consequences, serious consequences
postmodern, preconceived, psychoanalytic	key factors, key elements, main factors	contingencies, specific items, specific terms	features, oddities, standard features
military aid			
military	aid	military aid (observed)	military aid (estimated)
servicemen, commandos, military intelligence	guidance, advice, assistance	servicemen, commandos, military intelligence	guidance, advice, assistance
guerrilla paramilitary, anti-terrorist	mentoring, tutoring, internships	guidance, advice, assistance	servicemen, commandos, military intelligence
conglomerate, giants, conglomerates	award, awards, honors	compliments, congratulations, replies	mentoring, tutoring, internships
managerial, logistical, governmental	certificates, degrees, bachelor	training, appropriate training, advanced training	award, awards, honors
humankind, Palestinian people, Iraqi people	servicemen, commandos, military intelligence	conglomerate, giants, conglomerates	conglomerate, giants, conglomerates
bad behavior			
bad	behavior	bad behavior (observed)	bad behavior (estimated)
Great place, place, fantastic place	scholastic achievement, ethical behavior, behaviors	scholastic achievement, ethical behavior, behaviors	scholastic achievement, ethical behavior, behaviors
antithesis, affront, omen	dating, intimacy, courtship	intruders, intrusions, overflows	dating, intimacy, courtship
thankful, grateful, sorry	morphology, phylogeny, physiology	inconsistencies, faults, flaws	morphology, phylogeny, physiology
goofy, crazy, fucking	psychosis, depression, disorder	comm, wildness, haunting	psychosis, depression, disorder
go-ahead, spanking, shrift	invited, attitudes, encouraged	pasts, non-commercial use, mind-set	invited, attitudes, encouraged

Chapter 4

A Joint Model of Semantics in Corpus and the Brain

This work is published as "Interpretable Semantic Vectors from a Joint Model of Brain- and Text-Based Meaning" (Fyshe et al., 2014).

Vector space models (VSMs) of semantics represent word meanings as points in a high dimensional space. VSMs are typically created using a large text corpora, and so represent word semantics as observed in text. In this chapter, we present an algorithm (JNNSE) that can incorporate a measure of semantics not previously used to create VSMs: brain activation data recorded while people read words. The resulting model takes advantage of the complementary strengths and weaknesses of corpus and brain activation data to give a more complete representation of semantics. Evaluations show that, compared to a model that uses only one data source, our joint model 1) matches a behavioral measure of semantics more closely, 2) can be used to predict corpus data for unseen words and 3) has predictive power that generalizes across brain imaging technologies and across subjects. We believe that our model is thus a more faithful representation of mental vocabularies.

4.1 Introduction

Vector Space Models (VSMs) represent lexical meaning by assigning each word a point in high dimensional space. Beyond their use in computational linguistics applications, they are of interest to cognitive scientists as an objective and data-driven method to discover word meanings (Landauer and Dumais, 1997b).

Typically, VSMs are created by collecting word usage statistics from large amounts of text data and applying some dimensionality reduction technique like Singular Value Decomposition (SVD). The basic assumption is that semantics drives a person's language production behavior, and, as a result, co-occurrence patterns in written text indirectly encode word meaning. Raw co-occurrence statistics are unwieldy, but in the compressed VSM the distance between any two words is conceived to represent their mutual semantic similarity, as perceived and judged

by speakers (Sahlgren, 2006a; Turney and Pantel, 2010). This space then reflects the "semantic ground truth" of shared lexical meanings in a language community's vocabulary. However, corpus-based VSMs have been criticized as being noisy or incomplete representations of meaning (Glenberg and Robertson, 2000). For example, multiple word senses collide in the same vector, and noise from mis-parsed sentences or spam documents can interfere with the final semantic representation.

When a person is reading or writing, the semantic content of each word will be necessarily activated in the brain, and so in patterns of activity over individual neurons. In principle, brain activity could replace corpus data as input to a VSM, and contemporary brain imaging techniques allow us to attempt this. Functional Magnetic Resonance Imaging (fMRI) and Magnetoencephalography (MEG) are two brain activation recording technologies that measure neuronal activation in aggregate, and have been shown to have a predictive relationship with models of word meaning (Mitchell et al., 2008; Palatucci et al., 2009; Sudre et al., 2012; Murphy et al., 2012a). MEG measures the magnetic field caused by many neurons firing in a coordinated fashion; fMRI measures the change in blood oxygenation in response to change in neural activity. For a more complete discussion of brain imaging technologies, including examples of the data used in this Chapter, refer to Section 2.2.1.

If brain activation data encodes semantics, we theorized that including brain data in a model of semantics could result in a model more consistent with semantic ground truth. However, the inclusion of human brain data will only improve a text-based model if brain data contains semantic information not readily available in the corpus. In addition, if a semantic test involves another subject's brain activation data, performance can improve only if the additional semantic information is consistent across brains. Of course, brains differ in shape, size and in connectivity, so additional information encoded in one brain might not translate to another. Furthermore, different brain imaging technologies measure very different correlates of neuronal activity (see Section 2.2.1). Due to these differences, it is possible that one subject's brain activation data cannot improve a model's performance on another subject's brain data, or for brain data collected using a different recording technology. Indeed, inter-subject models of brain activation is an open research area (Conroy et al., 2013), as is learning the relationship between brain imaging technologies (Engell et al., 2012; Hall et al., 2013). Brain data can also be corrupted by many types of noise (e.g. recording room interference, movement artifacts), another possible hindrance to the use of brain data in VSMs.

VSMs are interesting from both engineering and scientific standpoints. This chapter focuses on the scientific questions: Can the inclusion of brain data improve semantic representations learned from corpus data? What can we learn from such a model? From an engineering perspective, brain activation data will likely never replace text data. Brain activation recordings are both expensive and time consuming to collect, whereas textual data is vast and much of it is free to download. However, from a scientific perspective, combining text and brain data could lead to more consistent semantic models, in turn leading to a better understanding of semantics and semantic modeling generally.

For this project, we leveraged both brain and text data to build a hybrid VSM using a new matrix factorization method (JNNSE). Our hypothesis is that the noise of brain and corpus derived statistics will be largely orthogonal, and so the two data sources will have complementary strengths as input to VSMs. If this hypothesis is correct, the resulting VSM should be more

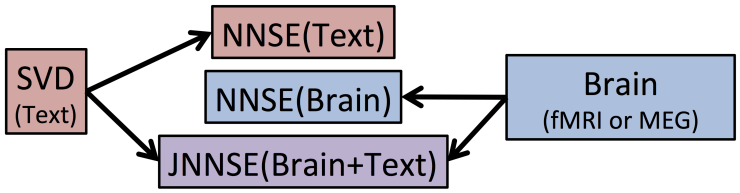


Figure 4.1: The models used in this chapter, along with their input data sources. SVD is the original corpus-based representation from Fyshe et al. (2013), fMRI data is from Mitchell et al. (2008), MEG data is from Sudre et al. (2012). NNSE models are built from only one data source, whereas JNNSE has the ability to join data sources.

successful in modeling word semantics as encoded in human judgements, and as encoded in separate corpus and brain data not used to create the model. In this chapter we will show that, when compared to a model that uses only text data, JNNSE:

1. creates a VSM that is more correlated to a behavioral measure of word semantics.
2. produces word vectors that are more predictable from the brain activity of different people, even when brain data is collected with a different recording technology.
3. predicts corpus representations of withheld words more accurately than a model that does not combine data sources.
4. directly maps semantic concepts onto the brain by learning a latent representation that maps between text and brain imaging data.

Together, these results suggest that corpus and brain activation data measure semantics in compatible and complementary ways. Our results are evidence that a joint model of brain- and text-based semantics may be closer to semantic ground truth than text-only models. Our findings also indicate that there is additional semantic information available in brain activation data that is not present in corpus data, implying that there are elements of semantics currently lacking in text-based VSMs. The top performing VSM created with brain and text data is available online (`http://www.cs.cmu.edu/~afyshe/papers/acl2014/`).

Figure 4.1 illustrates the different models covered in this chapter and the input data sources for each. In Sections 4.2 and 4.3 I will review NNSE, and our extension, JNNSE. In Section 4.4 will describe the data used, and in Section 4.5 the experiments that support our position that brain data is a valuable source of semantic information that compliments text data.

4.2 Non-Negative Sparse Embedding

As covered in Chapter 3, Non-Negative Sparse Embedding (NNSE) (Murphy et al., 2012b) is an algorithm that produces a latent representation using matrix factorization. Standard NNSE begins with a matrix $X \in \mathbb{R}^{w \times c}$ made of c corpus statistics for w words. NNSE solves the

following objective function:

$$\underset{A,D}{\operatorname{argmin}} \sum_{i=1}^{w} \left\| X_{i,:} - A_{i,:} \times D \right\|^2 + \lambda \left\| A \right\|_1 \tag{4.1}$$

$$\text{subject to: } D_{i,:} D_{i,:}^T \leq 1, \forall\, 1 \leq i \leq \ell \tag{4.2}$$

$$A_{i,j} \geq 0,\ 1 \leq i \leq w,\ 1 \leq j \leq \ell \tag{4.3}$$

The solution will find a matrix $A \in \mathbb{R}^{w \times \ell}$ that is sparse, non-negative, and represents word semantics in an ℓ-dimensional latent space. $D \in \mathbb{R}^{\ell \times c}$ gives the encoding of corpus statistics in the latent space. Together, they factor the original corpus statistics matrix X in a way that minimizes the reconstruction error. The L_1 constraint encourages sparsity in A; λ is a hyperparameter. Equation 4.2 constrains D to eliminate solutions where A is made arbitrarily small by making D arbitrarily large. Equation 4.3 ensures that A is non-negative. We may increase ℓ to give more dimensional space to represent word semantics, or decrease ℓ for more compact representations.

The sparse and non-negative representation in A produces a more interpretable semantic space, where interpretability is quantified with a behavioral task (Chang et al., 2009a; Murphy et al., 2012b). To illustrate the interpretability of NNSE, we describe a word by selecting the word's top scoring dimensions, and selecting the top scoring words in those dimensions. For example, the word chair has the following top scoring dimensions:

1. chairs, seating, couches;
2. mattress, futon, mattresses;
3. supervisor, coordinator, advisor.

These dimensions cover two of the distinct meanings of the word chair (furniture and person of power).

NNSE's sparsity constraint dictates that each word can have a non-zero score in only a few dimensions, which aligns well to previous feature elicitation experiments in psychology. In feature elicitation, participants are asked to name the characteristics (features) of an object. The number of characteristics named is usually small (McRae et al., 2005), which supports the requirement of sparsity in the learned latent space.

4.3 Joint Non-Negative Sparse Embedding

In Chapter 3 we extended NNSE to incorporate the notion of semantic composition. Here, we extend NNSEs to incorporate an additional source of data for a subset of the words in X, and call the approach Joint Non-Negative Sparse Embeddings (JNNSEs). The JNNSE algorithm is general enough to incorporate any new information about the a word w, but for this study we will focus on brain activation recordings of a human subject reading single words. We will incorporate either fMRI or MEG data, and call the resulting models JNNSE(fMRI+Text) and JNNSE(MEG+Text) and refer to them generally as JNNSE(Brain+Text). For clarity, from here on, we will refer to NNSE as NNSE(Text), or NNSE(Brain) depending on the single source of input data used.

Let us order the rows of the corpus data X so that the first $1 \dots w'$ rows have both corpus statistics and brain activation recordings. Each brain activation recording is a row in the brain data matrix $Y \in \mathbb{R}^{w' \times v}$ where v is the number of features derived from the recording. For MEG recordings, v =sensors $\times$ time points= 306×150. For fMRI v = grey-matter voxels $=\simeq 20,000$ depending on the brain anatomy of each individual subject. The new objective function is:

$$\begin{aligned}
\underset{A, D^{(c)}, D^{(b)}}{\operatorname{argmin}} \quad & \sum_{i=1}^{w} \left\| X_{i,:} - A_{i,:} \times D^{(c)} \right\|^2 + \\
& \sum_{i=1}^{w'} \left\| Y_{i,:} - A_{i,:} \times D^{(b)} \right\|^2 + \lambda \left\| A \right\|_1 && (4.4) \\
\text{subject to:} \quad & D^{(c)}_{i,:} D^{(c)^T}_{i,:} \leq 1, \forall\, 1 \leq i \leq \ell && (4.5) \\
& D^{(b)}_{i,:} D^{(b)^T}_{i,:} \leq 1, \forall\, 1 \leq i \leq \ell && (4.6) \\
& A_{i,j} \geq 0,\ 1 \leq i \leq w,\ 1 \leq j \leq \ell && (4.7)
\end{aligned}$$

We have introduced an additional constraint on the rows $1 \dots w'$, requiring that some of the learned representations in A also reconstruct the brain activation recordings (Y) through representations in $D^{(b)} \in \mathbb{R}^{\ell \times v}$. Let us use A' to refer to the brain-constrained rows of A ($A' \subseteq A$). Words that are close in "brain space" must have similar representations in A', which can further percolate to affect the representations of other words in A via closeness in "corpus space".

With two of $A, D^{(c)}$ or $D^{(b)}$ fixed, the objective function for NNSE(Text) and JNNSE(Brain+Text) is convex. However, we are solving for $A, D^{(c)}$ and $D^{(b)}$ simultaneously, so the problem is non-convex. To solve for this objective, we use the online algorithm of Section 3 from Mairal et al. (2010). This algorithm is guaranteed to converge, and in practice we found that JNNSE(Brain+Text) converged as quickly as NNSE(Text) for the same ℓ. This algorithm was an easy extension to NNSE(Text) and required very little additional tuning. To solve for $D^{(c)}$ and $D^{(b)}$ we use the same gradient descent method as in Chapter 3: . When solving for A, the system of constraints in Equation 4.4 simplifies to lasso regression, for which we use the SPAMS package[1] to solve, and set $\lambda = 0.025$.

We also consider learning shared representations in the case where data X and Y contain the effects of known *disjoint* features. For example, when a person reads a word, the recorded brain activation data Y will contain the neural activity associated with perceiving the stimulus, which is unrelated to the semantics of the word. These signals can be attributed to, for example, the number of letters in the word and the number of white pixels on the screen (Sudre et al., 2012). To account for such effects in the data, we augment A' with a set of $n = 11$ fixed, manually

[1]SPAMS Package: `http://spams-devel.gforge.inria.fr/`

defined features (e.g. word length) to create $A_{percept} \in \mathbb{R}^{w \times (\ell+n)}$. The optimization becomes:

$$\begin{aligned}
\underset{A, D^{(c)}, D^{(b)}}{\operatorname{argmin}} \quad & \sum_{i=1}^{w} \left\| X_{i,:} - A_{i,:} \times D^{(c)} \right\|^2 + \\
& \sum_{i=1}^{w'} \left\| Y_{i,:} - [A_{i,:} | A_{percept}] \times D^{(b)} \right\|^2 + \lambda \left\| A \right\|_1 && (4.8) \\
\text{subject to:} \quad & D^{(c)}_{i,:} D^{(c)^T}_{i,:} \leq 1, \forall\, 1 \leq i \leq \ell && (4.9) \\
& D^{(b)}_{i,:} D^{(b)^T}_{i,:} \leq 1, \forall\, 1 \leq i \leq \ell && (4.10) \\
& A_{i,j} \geq 0,\ 1 \leq i \leq w,\ 1 \leq j \leq \ell && (4.11)
\end{aligned}$$

Where $[X|Y]$ indicates the concatenation of matrices X and Y. The extended $D^{(b)} \in \mathbb{R}^{(\ell+n) \times v}$ is used with A' and $A'_{percept}$, to reconstruct the brain data Y. More generally, one could instead allocate a certain number of latent features specific to X *or* Y, both of which could be learned, as explored in some related work (Gupta et al., 2013). We use 11 *perceptual* features that characterize the non-semantic features of the word stimulus (for a list, see Sudre et al. (2012) or `http://www.cs.cmu.edu/~fmri/neuroimage2012_files/features.html`).

The JNNSE algorithm is advantageous in that it can handle partially paired data. That is, the algorithm does not require that every row in X also have a row in Y. Fully paired data is a requirement of many other approaches (White et al., 2012; Jia and Darrell, 2010). Our approach allows us to leverage the semantic information in corpus data even for words without brain activation recordings.

JNNSE(Brain+Text) does not require brain data to be mapped to a common average brain, a step that is often necessary when one wants to generalize brain activity between human subjects. Such mappings can blur and distort data, making it less useful for subsequent prediction steps. We avoid these mappings, and instead use the fact that similar words elicit similar brain activation *within* a subject. In the JNNSE algorithm, it is this closeness in "brain space" that guides the creation of the latent space A. Leveraging intra-subject distance measures to study inter-subject encodings has been studied previously (Kriegeskorte et al., 2008a; Raizada and Connolly, 2012), and has even been used across species (humans and primates) (Kriegeskorte et al., 2008b).

Though we restrict ourselves to using one subject per JNNSE(Brain+Text) model, the JNNSE algorithm could easily be extended to include data from multiple brain imaging experiments by adding a new squared loss term for additional brain data.

4.3.1 Related Work

Perhaps the most well known related approach to joining data sources is Canonical Correlation Analysis (CCA) (Hotelling, 1936), which has been applied to brain activation data in the past (Rustandi et al., 2009). CCA seeks two linear transformations that maximally correlate two data sets (X and Y) in the transformed forms dictated by linear mappings a and b.

$$\underset{a,b}{\operatorname{argmax}} \quad \operatorname{corr}(aX, bY) \tag{4.12}$$

where corr is correlation. CCA requires that the data sources be paired (all rows in the corpus data must have a corresponding brain data), as correlation between points is integral to the objective and requires paired data. To apply CCA to our data we would need to discard the vast majority of our corpus data, and use only the 60 rows of X with corresponding rows in Y. Compare the loss function in Equation 4.12 to that in Equation 4.4. Note that, while CCA holds the input data fixed and maximally correlates the transformed form, JNNSE holds the transformed form (A) fixed and seek a solution that maximally correlates the reconstruction ($AD^{(c)}$ or $A'D^{(b)}$) with the input data (X and Y respectively). This shift in error compensation is what allows our data to be only partially paired, and also results in a single learned representation that is shared between both data inputs.. While a Bayesian formulation of CCA can handle missing data, our model has missing data for $> 97\%$ of the full $w \times (v + c)$ brain and corpus data matrix. To our knowledge, this extreme amount of missing data has not been explored with Bayesian CCA.

One could also use a topic model style formulation to represent this semantic representation task. Supervised topic models (Blei and McAuliffe, 2007) use a latent topic to generate two observed outputs: words in a document and a categorical label for the document. The same idea could be applied here: the latent semantic representation generates the observed brain activity and corpus statistics. Generative and discriminative models both have their own strengths and weaknesses, generative models being particularly strong when data sources are limited (Ng and Jordan, 2002). Our task is an interesting blend of data-limited and data-rich problem scenarios.

In the past, various pieces of additional information have been incorporated into semantic models. For example, models with behavioral data (Silberer and Lapata, 2012) and models with visual information (Bruni et al., 2011; Silberer et al., 2013) have both shown to improve semantic representations. Other works have correlated VSMs built with text or images with brain activation data (Murphy et al., 2012a; Anderson et al., 2013). To our knowledge, this work is the first to integrate brain activation data into the construction of the VSM.

4.4 Data

4.4.1 Corpus Data

The corpus statistics used here are the vectors from Fyshe et al. (2013). They are compiled from a 16 billion word subset of ClueWeb09 (Callan and Hoy, 2009) and contain two types of corpus features: dependency and document features, found to be complementary for most tasks. Dependency statistics were derived by dependency parsing the corpus and compiling counts for all dependencies incident on the word. Document statistics are word-document co-occurrence counts. Count thresholding was applied to reduce noise, and positive pointwise-mutual-information (PPMI) (Church and Hanks, 1990) was applied to the counts. SVD was applied to the document and dependency statistics and the top 1000 dimensions of each type were retained. We selected the rows corresponding to noun-tagged words (approx. 17000 words). Throughout this chapter this dataset will be referred to as SVD, and is the text data input to JNNSE and NNSE models.

4.4.2 Brain Activation Data

We have MEG from Sudre et al. (2012) and fMRI data from Mitchell et al. (2008) at our disposal. The fMRI data and MEG data are from 18 subjects (9 in each imaging modality) viewing 60 concrete nouns (Mitchell et al., 2008; Sudre et al., 2012). The 60 words span 12 word categories (animals, buildings, tools, insects, body parts, furniture, building parts, utensils, vehicles, objects, clothing, food). Each of the 60 words was presented with a line drawing, so word ambiguity is not an issue. For both recording modalities, all trials for a particular word were averaged together to create one training instance per word, with 60 training instances in all for each subject and imaging modality. Preprocessing details for the MEG data appears in Sudre et al. (2012) and for fMRI data in Mitchell et al. (2008).

Table 4.1: A Comparison of the models explored in this chapter, and the data upon which they operate.

Model Name	Section(s)	Text Data	Brain Data	Withheld Data
NNSE(Text)	4.2, 4.5	✓	x	-
NNSE(Brain)	4.2, 4.5.2, 4.5.3	x	✓	-
JNNSE(Brain+Text)	4.3, 4.5	✓	✓	-
JNNSE(Brain+Text): Dropout task	4.5.2	✓	✓	subset of brain data
JNNSE(Brain+Text): Predict corpus	4.5.3	✓	✓	subset of text data

4.5 Experimental Results

Here we explore several variations of JNNSE and NNSE formulations. For a comparison of the models used, see Table 4.1.

4.5.1 Correlation to Behavioral Data

To test if our joint model of Brain+Text is closer to semantic ground truth than a model that uses only text data, we compared the latent representation A learned via JNNSE(Brain+Text) or NNSE(Text) to an independent behavioral measure of semantics. We collected behavioral data for the 60 nouns in the form of answers to 218 semantic questions. Answers were gathered with Mechanical Turk. The full list of questions appears in Sudre et al. (2012). Some example questions are:"Is it alive?", and "Can it bend?". Mechanical Turk users were asked to respond to each question for each word on a scale of 1-5. At least 3 users answered each question and the median score was used. This gives us a semantic representation of each of the 60 words in a 218-dimensional behavioral space. Answers were required for each questions, thus we do not have the problems of sparsity that exist for feature production norms from other studies (McRae et al., 2005). In addition, the answers are ratings, rather than binary yes/no answers.

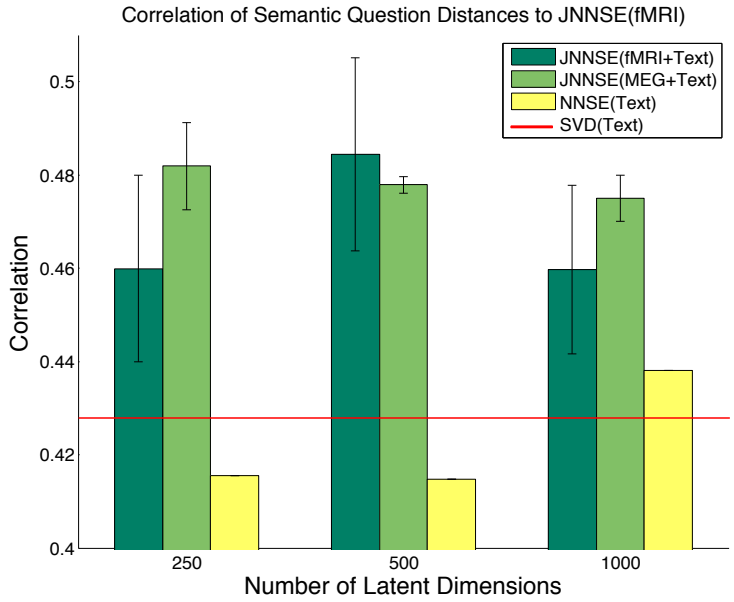


Figure 4.2: The correlation of pairwise word distances from several models to the pairwise word distances based on behavioral data. Error bars indicate SEM.

For a given value of ℓ we solve the NNSE(Text) and JNNSE(Brain+Text) objective function as detailed in Equation 4.1 and 4.4 respectively. We compared JNNSE(Brain+Text) and NNSE(Text) models by first computing the pairwise distances between all words in each learned latent space. We then measure the correlation of those distances to the pairwise distances in the 218-dimensional semantic space. Distances were calculated using normalized Euclidean distance (equivalent in rank-ordering to cosine distance, but more suitable for sparse vectors). Figure 4.2 shows the results of this correlation test. The error bars for the JNNSE(Brain+Text) models represent a 95% confidence interval calculated using the standard error of the mean (SEM) over the 9 person-specific JNNSE(Brain+ Text) models. Because there is only one NNSE(Text) model for each dimension setting, no SEM can be calculated, but it suffices to show that the NNSE(Text) correlation does not fall into the 95% confidence interval of the JNNSE(Brain+Text) models. The SVD matrix for the original corpus data has correlation 0.4279 to the behavioral data, also below the 95% confidence interval for all JNNSE models. The results show that a model that incorporates brain activation data is more faithful to a behavioral measure of semantics.

4.5.2 Word Prediction from Brain Activation

Compared to NNSE models, JNNSE(Brain+Text) vectors allow us to do a better job of predicting the word a different person is reading. This increased predictability implies that a learned representation that incorporates one subject's brain data is more consistent with other independent samples of brain activity collected from different people. Because there is a large degree of variation between brains and because MEG and fMRI measure very different correlates of neuronal activity, this type of generalization has proven to be very challenging and is an open research question in the neuroscience community (Conroy et al., 2013).

The output A of the JNNSE(Brain+Text) or NNSE(Text) algorithm can be used as a VSM, which we use for the task of word prediction from fMRI or MEG recordings. JNNSE(Brain+Text) vectors created with a particular human subject's data is never used in the prediction framework with that same subject. For example, if we use fMRI data from subject 1 to create a JNNSE(fMRI+Text), we will test it with the remaining 8 fMRI subjects, but all 9 MEG subjects (fMRI and MEG subjects are disjoint).

Let us call the VSM learned with JNNSE(Brain+Text) or NNSE(Text) the *semantic vectors*. We can train a weight matrix (W) that predicts the semantic vector ($\mathbf{a}$) of a word from that word's brain activation vector $\mathbf{x}$: $\mathbf{a} = W\mathbf{x}$. W can be learned with a variety of methods, we will use L_2 regularized regression. One can also train regressors that predict the brain activation data from the semantic vector: $\mathbf{x} = W\mathbf{a}$, but we have found this to give lower predictive accuracy. Note that we must *re-train* our weight matrix W for each subject (instead of re-using $D^{(b)}$ from Equation 4.4) because testing always occurs on a different subject, and the brain activation data is not inter-subject aligned.

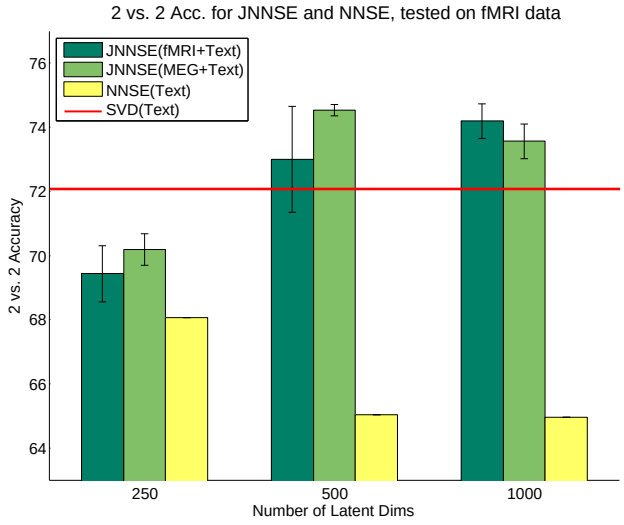


Figure 4.3: Average 2 vs. 2 accuracy for predicting words from fMRI data. Performance is calculated using semantic vectors from SVD, NNSE(Text) or JNNSE(Brain+Text). A different JNNSE(Brain+Text) is trained for each subject's brain data; the brain image data used to create a model is never used to test that same model. For JNNSE(Brain+Text), error bars show SEM calculated across the subject specific models.

We train ℓ independent L_2 regularized regressors to predict the ℓ-dimensional vectors $\mathbf{a} = \{a_1 \dots a_\ell\}$. The predictions are concatenated to produce a *predicted* semantic vector: $\hat{\mathbf{a}} = \{\hat{a}_1, \dots, \hat{a}_\ell\}$. We assess word prediction performance by testing if the model can differentiate between two words, a task named *2 vs. 2 prediction* (Mitchell et al., 2008; Sudre et al., 2012). We choose the assignment of the two held out semantic vectors $(\mathbf{a}^{(1)}, \mathbf{a}^{(2)})$ to predicted semantic vectors $(\hat{\mathbf{a}}^{(1)}, \hat{\mathbf{a}}^{(2)})$ that minimizes the sum of the two normalized Euclidean distances. 2 vs. 2 accuracy is the percentage of tests where the correct assignment is chosen.

The 60 nouns fall into 12 word categories. Words in the same word category (e.g. screwdriver

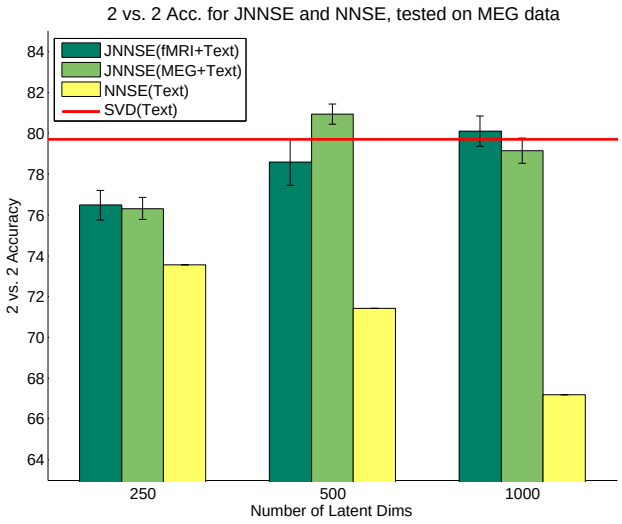


Figure 4.4: Average 2 vs. 2 accuracy for predicting words from MEG data. Performance is calculated using semantic vectors from SVD, NNSE(Text) or JNNSE(Brain+Text). A different JNNSE(Brain+Text) is trained for each subject's brain data; the brain image data used to create a model is never used to test that same model. For JNNSE(Brain+Text), error bars show SEM calculated across the subject specific models.

and hammer) are closer in semantic space than words in different word categories, which makes some 2 vs. 2 tests more difficult than others. We choose 150 random pairs of words (with each word represented equally) to estimate the difficulty of a typical word pair, without having to test all $\binom{60}{2}$ word pairs. The same 150 random pairs are used for all subjects and all VSMs. Expected chance performance on the 2 vs. 2 test is 50%.

Results for testing on fMRI data in the 2 vs. 2 framework appear in Figure 4.3. JNNSE(fMRI+ Text) data performed on average 6% better than the best NNSE(Text), and exceeding even the original SVD corpus representations, while maintaining interpretability. 95% confidence intervals calculated using the standard error of the mean (SEM) show no significant difference between JNNSE(MEG+Text) or JNNSE(fMRI+Text), but a significant advantage over NNSE and the original SVD models. Our results generalize across brain activity recording types; JNNSE(MEG+ Text) performs as well as JNNSE(fMRI+ Text) when tested on fMRI data. The results are consistent when testing on MEG data: JNNSE(MEG+ Text) or JNNSE(fMRI+ Text) outperforms NNSE(Text) (see Figure 4.4), however, only the 500-dimension JNNSE(MEG+Text) model outperforms the SVD baseline. Still, JNNSE models improve interpretability over SVD models, and a joint model is still advantageous when testing on MEG data, even if it does not quantitatively outperform an SVD model.

NNSE(Text) performance decreases as the number of latent dimension increases. This implies that without the regularizing effect of brain activation data, the extra NNSE(Text) dimensions are being used to overfit to the corpus data, or possibly to fit semantic properties not detectable with current brain imaging technologies. However, when brain activation data is included, increasing the number of latent dimensions strictly increases performance for JNNSE

(fMRI+Text). JNNSE(MEG+Text) has peak performance with 500 latent dimensions, with $\sim 1\%$ decrease in performance at 1000 latent dimensions. In previous work, the ability to decode words from brain activation data was found to improve with added latent dimensions (Murphy et al., 2012b). Our results may differ because our words are POS tagged, and we included only nouns for the final NNSE(Text) model. We found that with the original $\lambda = 0.05$ setting from Murphy et al. (2012b) produced vectors that were too sparse; four of the 60 words had all-zero vectors, making differentiation amongst those words during the 2 vs. 2 test impossible. When $\lambda = 0.05$, JNNSE(Brain+Text) did not have any all-zero vectors. To improve the NNSE(Text) vectors for a fair comparison, we reduced $\lambda = 0.025$, under which NNSE(Text) did not produce any all-zero vectors for the 60 words.

Our results show that brain activation data contributes additional information, over and above the information available in corpus data. This leads to an increase in performance for the task of word prediction from brain activation data. This suggests that current corpus-only models may not capture all relevant semantic information. This conflicts with previous studies which found that semantic vectors culled from corpus statistics contain all of the semantic information required to predict brain activation (Bullinaria and Levy, 2013).

Prediction from a Brain-only Model

How much predictive power does the corpus data provide to this word prediction task? To test this, we calculated the 2 vs. 2 accuracy for a NNSE(Brain) model trained on brain activation data only. We train NNSE(Brain) with one subject's data and use the resulting vectors to calculate 2 vs. 2 accuracy for the remaining subjects. We have brain data for only 60 words, so using $\ell \geq 60$ latent dimensions leads to an under-constrained system and a degenerate solution wherein only one latent dimension is active for any word (and where the brain data can be perfectly reconstructed). The degenerate solution makes it impossible to generalize across words and leads to performance at chance levels. An NNSE(MEG) trained on MEG data gave maximum 2 vs. 2 accuracy of 67% when $\ell = 20$. The reduced performance may be due to the limited training data and the low SNR of the data, but could also be attributed to the lack of corpus information, which provides another piece of semantic information.

Effect on Rows Without Brain Data

It is possible that some JNNSE(Brain+Text) dimensions are being used exclusively to fit brain activation data, and not the semantics represented in both brain and corpus data. If a particular dimension j is solely used for brain data, the sparsity constraint will favor solutions that sets $A_{(i,j)} = 0$ for $i > w'$ (no brain data constraint), and $A_{(i,j)} > 0$ for some $0 \leq i \leq w'$ (brain data constrained). We found that there were no such dimensions in the JNNSE(Brain+Text). In fact for the $\ell = 1000$ JNNSE(Brain+Text), all latent dimensions had greater than $\sim 25\%$ non-zero entries, which implies that all dimensions are being shared between the two data inputs (corpus and brain activation), and are used to reconstruct both.

To test that the brain activation data is truly influencing rows of A not constrained by brain activation data, we performed a *dropout* test. Again, we split the original 60 words into two 30 word groups (as evenly as possible across word categories). We trained JNNSE(fMRI+Text)

with 30 words, simulating the scenario where we have corpus data but no brain data for some words and some subjects. We then use the rows of the resulting A matrix corresponding to the 30 withheld words to test 2 vs. 2 accuracy with the remaining 8 fMRI subjects. The testing data has been halved, so we used 75 randomly chosen word pairs instead of 150. Because the words are disjoint, we could have tested with all 9 subjects, but for the most accurate comparison, we followed the same methodology as the previous 2 vs. 2 test.

Along with the results from the original 2 vs. 2 results for all 60 words, Figure 4.5 shows the results for the dropout scenario for JNNSE(fMRI+Text) with $\ell = 1000$ latent dimensions tested on fMRI data. In this scenario, each time we perform a 2 vs. 2 test we are training on 28 instead of 58 words, and so we expect performance to suffer. For NNSE(Text), performance on the 2 vs. 2 task with only 30 words is very low, 55.6%. The drop in accuracy is due only to the reduction in training data, as there is no change in the semantic vectors used to perform the 2 vs. 2 test. The results are lower for JNNSE(fMRI+Text) tested on 30 words, but is still 7% higher than results with NNSE(Text). Because the training and testing words are completely disjoint, these results imply that the addition of brain activation data improves the learned latent representation, not only for those words for which we have brain activation data, but also for the words for which there is no brain activation data. This, along with the fact that all latent dimensions used by words with brain data are also used by words without brain data, suggests that our algorithm produces semantic representations that are better constrained for all words in A, even though we only add explicit additional constraints to a small number of words. The dropout test also shows that we could have collected a different set of 60 word for each of the 9 subjects for a total of 540 words and still successfully used JNNSE(Brain+Text). In the future, this insight will allow us to increase our coverage over words, which could lead to greatly improved semantic models.

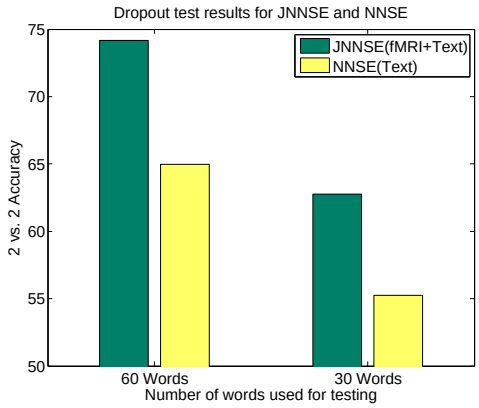


Figure 4.5: Performance on the dropout test (excluding 30 words of input brain data), as tested on fMRI. We compare JNNSE(fMRI+Text) with $\ell = 1000$ when train/tested on the same 60 words (same as rightmost bars in Figure 4.3) and when train/tested on disjoint sets of 30 words. Performance decreases for both JNNSE(fMRI+Text) and NNSE(Text), but JNNSE(fMRI+Text) still outperforms NNSE(Text).

4.5.3 Predicting Corpus Data

Can an accurate latent representation of a word be constructed using only brain activation data? This task simulates the scenario where there is no reliable corpus representation of a word, but brain data is available. This scenario may occur for seldom-used words that fall below the thresholds used for the compilation of corpus statistics. It could also be useful for acronym tokens (lol, omg) found in social media contexts where the meaning of the token is actually a full sentence.

We trained a JNNSE(fMRI+Text) with brain data for all 60 words, but withhold the corpus data for 30 of the 60 words (as evenly distributed as possible amongst the 12 word categories). The brain activation data for the 30 withheld words will allow us to create latent representations in A for withheld words. Simultaneously, we will learn a mapping from the latent representation to the corpus data ($D^{(c)}$). Let $w^{(c)}$ be the set of words with corpus data, let $w^{(b)}$ be the words with brain data. In previous experiments, $w^{(b)}$ was a subset of $w^{(c)}$, but here the two sets overlap by only 30 elements, and there are 30 elements in $w^{(b)}$ that have no corresponding element in $w^{(c)}$. The summation in the optimization is now over these two sets:

$$\begin{aligned}
\underset{A,D^{(c)},D^{(b)}}{\operatorname{argmin}} & \sum_{i \in w^{(c)}} \left\| X_{i,:} - A_{i,:} \times D^{(c)} \right\|^2 + \\
& \sum_{i \in w^{(b)}} \left\| Y_{i,:} - [A_{i,:} | A_{percept}] \times D^{(b)} \right\|^2 + \lambda \left\| A \right\|_1 && (4.13) \\
\text{subject to:} \quad & D^{(c)}_{i,:} D^{(c)^T}_{i,:} \leq 1, \forall\, 1 \leq i \leq \ell && (4.14) \\
& D^{(b)}_{i,:} D^{(b)^T}_{i,:} \leq 1, \forall\, 1 \leq i \leq \ell && (4.15) \\
& A_{i,j} \geq 0,\ 1 \leq i \leq w,\ 1 \leq j \leq \ell && (4.16)
\end{aligned}$$

This task cannot be performed with a NNSE(Text) model because one cannot learn a latent representation of a word without data of some kind. This further emphasizes the impact of brain imaging data, which will allow us to generalize to previously unseen words in corpus space.

We use the latent representations in A for each of the words without corpus data and the mapping to corpus space $D^{(c)}$ to predict the withheld corpus data in X. We then rank the withheld rows of X by their distance to the predicted row of X and calculate the mean rank accuracy of the held out words. Results in Table 4.2 show that we can recreate the withheld corpus data using brain activation data. Peak mean rank accuracy (67.37) is attained at $\ell = 500$ latent dimensions. This result shows that neural semantic representations can create a latent representation that is faithful to unseen corpus statistics, providing further evidence that the two data sources share a strong common element.

How much power is the remaining corpus data supplying in scenarios where we withhold corpus data? To answer this question, we trained an NNSE(Brain) model on 30 words of brain activation, and then trained a regressor to predict corpus data from those latent brain-only representations. We use the trained regressor to predict the corpus data for the remaining 30 words. Peak performance is attained at $\ell = 10$ latent dimensions, giving mean rank accuracy of 62.37, significantly worse than the model that includes both corpus and brain activation data (67.37).

Table 4.2: Mean rank accuracy over 30 words using corpus representations predicted by a JNNSE(MEG+Text) model trained with some rows of the corpus data withheld. Significance is calculated using Fisher's method to combine p-values for each of the subject-dependent models.

Latent Dim size	Rank Accuracy	p-value
250	65.30	$< 10^{-19}$
500	67.37	$< 10^{-24}$
1000	63.47	$< 10^{-15}$

4.5.4 Mapping Semantics onto the Brain

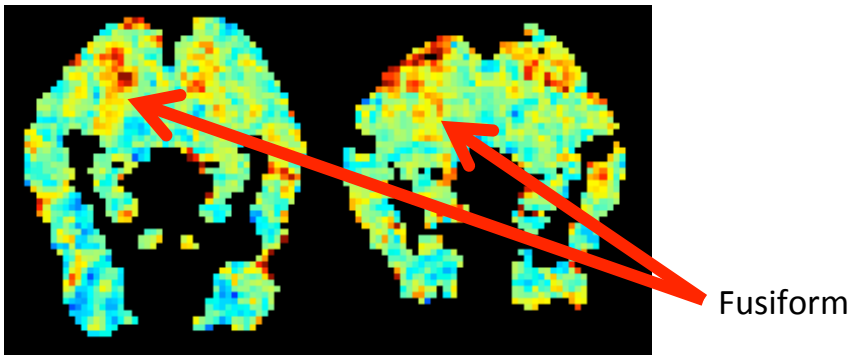
Because our method incorporates brain data into an interpretable semantic model, we can directly map semantic concepts onto the brain. To do this, we examined the mappings from the latent space to the brain space via $D^{(b)}$. We found that the most interpretable mappings come from models where the perceptual features had been scaled down (divided by a constant factor), which encourages more of the data to be explained by the semantic features in A. Figure 4.6 shows the mappings ($D^{(b)}$) for dimensions related to shelter, food and body parts. The red areas align with areas of the brain previously known to be activated by the corresponding concepts (Mitchell et al., 2008; Just et al., 2010). Our model has learned these mappings in an unsupervised setting by relating semantic knowledge gleaned from word usage to patterns of activation in the brain. This illustrates how the interpretability of JNNSE can allow one to explore semantics in the human brain. All mappings for one subject can be viewed at (`http://www.cs.cmu.edu/~afyshe/papers/acl2014/`).

4.6 Conclusion

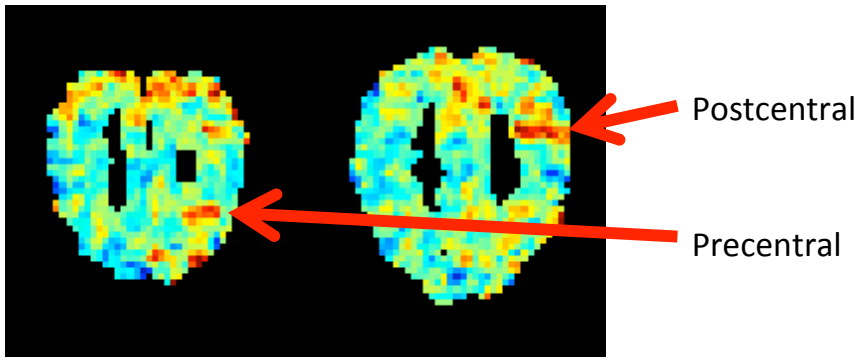
In this chapter, we presented an algorithm (JNNSE) that incorporates a novel measure of semantics: brain activation data recorded while people read words. Though the number of words for which we have brain image data is comparatively small, we have shown that including even this small amount of data has a positive impact on the learned latent representations. By comparing to a model that uses only one input data source, we measured this positive impact in several ways.

We showed that JNNSE representations are more correlated to a behavioral measure of semantics, can predict corpus representations of held out words more accurately, and can be used to more accurately predict the word a different person is reading based on their neural activity, even when that neural activity is recorded with a different brain imaging technology. We showed that this positive impact can percolate through to improve the representations of words for which we have no brain image data.

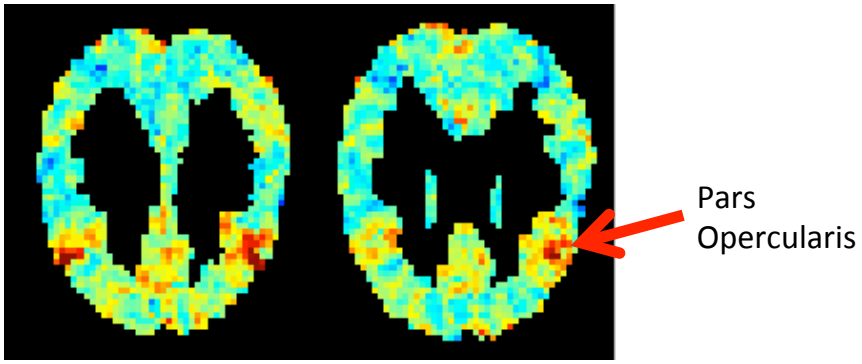
Previous work has shown that text-based models of semantics are consistent with neural activity (Mitchell et al., 2008; Murphy et al., 2012a). The results of this chapter reveal that there are aspects of neural semantic representations that are not fully represented in text-only VSMs. Our experiments show that brain- and corpus-based representations of semantics are both consistent and complementary. Our findings also indicate that we can use the brain as a semantic test, and



(a) $D^{(b)}$ matrix, subject P3, dimension with top words bathroom, balcony, kitchen. MNI coordinates z=-12 (left) and z=-18 (right). Fusiform is associated with shelter words.



(b) $D^{(b)}$ matrix; subject P1; dimension with top words ankle, elbow, knee. MNI coordinates z=60 (left) and z=54 (right). Pre- and post-central areas are activated for body part words.



(c) $D^{(b)}$ matrix; subject P1; dimension with top scoring words buffet, brunch, lunch. MNI coordinates z=30 (left) and z=24 (right). Pars opercularis is believed to be part of the gustatory cortex, which responds to food related words.

Figure 4.6: The mappings ($D^{(b)}$) from latent semantic space (A) to brain space (Y) for fMRI and words from three semantic categories. Shown are representations of the fMRI slices such that the back of the head is at the top of the image, the front of the head is at the bottom.

that such a test can differentiate models that capture this additional neurosemantic information from those that do not.

We have provided evidence that, when compared to a model that utilizes only one data source, the latent representations learned by our joint model are closer to the neural semantic representation of a different human subject, and to behavioral measurements of semantics averaged over several human subjects. This evidence leads us to believe that JNNSE representations may be closer to the "semantic ground truth" of shared lexical meanings that form a language community's vocabulary.

Chapter 5

Semantic Composition in the Brain

As a person reads, their brain performs a complex set of operations. Stimulus words are perceived, their semantics retrieved, ambiguity is resolved, individual word semantics are combined, and some final representation is created. While these steps are performed effortlessly by competent readers, relatively little is know about how the brain performs these actions, and where the final composed semantic representation is held in the brain. In this chapter, we explore semantic composition by analyzing Magnetoencephalography (MEG) recordings of the brain's activity as a person reads adjective-noun phrases. We will explore how the neural representation of the adjective, noun and phrase unfold over time and over areas of the brain. From these patterns we formulate a theory for adjective noun composition in the brain.

5.1 Introduction

Semantic composition is the process of combining small linguistic units to build more complex meaning, and is one of the more fundamental cognitive tasks required for language comprehension. It is a skill that children acquire with amazing speed, and that adults perform with little effort. Still, very little is known about the brain processes involved in semantic composition, and the neural representation of composed meaning.

In this chapter we study semantic composition in the human brain. As discussed in Section 2.2, composition in the brain has been studied previously in semantically anomalous sentences (Kutas and Hillyard, 1980; Kuperberg, 2007), as well as in simple phrases (Bemis and Pylkkänen, 2011; Baron, 2012). To our knowledge, the study presented here represents the first time that the *final phrasal meaning* of the semantic composition of adjective noun phrases has been studied with the fine time resolution offered by Magnetoencephalography (MEG).

Throughout this chapter we will test decodability from neural activity. Decodability is the ability to predict properties of the input stimuli from recordings of neural activity, and is a direct result of the information present in neural activity. We will refer also to encoding: the pattern of neural activity that represents a property (or properties) of the input stimuli. With this definition of encoding and decodability, when in time and where in brain-space we can decode a particular property is indicative of when and where information is encoded neurally. Note that the implication flows only in one direction, we do not claim that our inability to decode a property at a

particular time/space point is proof that the property is not encoded there. Lack of decodability could be due to many factors, including noise and the limitations of brain imaging technology, not necessarily to a property of the brain. In addition, what we call the encoding of a stimuli property could actually be the neural encoding of some correlated property. Even with these caveats, decodability is a useful concept for exploring neural information processing.

In this chapter we will use several decoding tasks to trace information flow in the brain, both at the word level and at the phrase level. We will trace the flow of information using several computational decoding tasks that vary in the amount and type of semantic information they are designed to detect. We will use the accuracy of decoding to infer where and when in the brain the processes of semantic composition occur and what their timing and locality imply about how the brain encodes and combines the meaning of word units.

This chapter begins with an outline of the decoding tasks of interest (Sections 5.1.1), and the data we collected to explore those tasks (Section 5.2). We then describe the framework used to calculate decodability, and methods used to determine chance levels of decodability (Section 5.3). We relay the results for each of the decoding tasks in Sections 5.4-5.6 and discuss these results in context in Section 5.7. We finish with a theory of adjective noun composition in the brain and conclusion in Sections 5.8 and 5.9.

5.1.1 Decoding Tasks

The following decoding tasks analyze different aspects of the meaning of adjective noun phrases. In each case, the task is to predict some property of the stimulus from the MEG recording. Differences in the time course of decodability for each task will help us infer the order in which the brain processes information during adjective noun phrase comprehension.

Adjective Attribute The adjective attribute refers to the property of the noun to be modulated by the adjective. For example, "big" modulates the size of a noun, "happy" modulates mood, and "yellow" modulates color. If we had several color adjectives in our experiment, they would all be instances of adjectives which modulate the attribute "color", and so would all belong to the same adjective attribute class. We are interested in how the decodability of the attribute being modulated differs from the decodability of the semantics of the adjective itself.

Adjective Semantics Adjective semantics are unique to each adjective, and are a cross between the attribute modulated, and the specific modulation that is occurring. For example, "big" is the modulation of size in the positive direction by some subjective amount, "yellow" is the color perceived in light with wavelength 570 nm, "happy" is a positive mood state. We used a large text corpus to automatically derive features, which are a proxy to adjective semantics. For example, the word "happy" is more often used to describe animate objects, and so that pattern of usage will make it more similar to words like "sad" than words that can apply to both animate and inanimate objects, like "red". More details on using a corpus as a proxy to semantics appears in Section 2.1.1

Noun Semantics Noun semantics are the unique semantic representation of a concept or object. In general, noun semantics can be thought of as the instantiation of several attributes with particular values. For example, banana is a noun with attribute value pairs edible=true, color=yellow, has_peel=true, shape=oblong, length=7 inches, sweetness=8, etc. When we decode the semantics of a noun we are decoding neural correlates of all of these attribute value assignments. We used a large text corpus to automatically derive features which are a proxy to noun semantics. For example the word "apple" might be more often used with verbs "eat" and "bite", which will make it similar to other food nouns.

Phrasal Semantics Phrasal semantics are the semantic representation (i.e. attribute value pairing) of a noun phrase. Since nouns themselves are noun phrases, the more general noun phrase occupies the same semantic space as nouns. In the case of adjective noun phrases, the semantics of the noun phrase is affected by the semantics of the adjective, the noun and the interaction of the noun and adjective. For example, the phrase "green banana" would have all of the same semantics of a banana, except that color=green, and sweetness=6. This illustrates how an adjective can modulate a particular attribute value explicitly (here: color) but also have implications for other semantic attributes (sweetness). We collected behavioral data to estimate some of the attribute value pairs for each adjective noun phrase.

5.2 Experimental Paradigm

Due to the rapid processing of phrasal semantics (most adults can read at a rate of 2-3 words per second) Magnetoencephalography is our brain imaging modality of choice. A full explanation of brain imaging technologies is given in Section 2.2.1. As a brief reminder, Magnetoencephalography (MEG) is a brain imaging technology that measures the magnetic field caused by many neurons firing in a coordinated fashion. MEG has better time resolution (as high as 1000 Hz) with slightly poorer spatial resolution when compared to fMRI.

To study the neural encoding of adjective, noun and phrasal semantics, we visually presented adjective noun phrases in isolation and recorded neural activity of 9 human subjects via a 306 channel Elekta Neuromag MEG machine. We chose nouns for this study that have been shown to be easily decodable (Sudre et al., 2012). Adjectives were chosen to modulate the most decodable semantic qualities of the words (e.g. edibility, manipulability, size).

For this study we selected 8 adjectives ("big", "small", "ferocious", "gentle", "light", "heavy", "rotten", ' 'tasty") and 6 nouns ("dog", "bear", "tomato", "carrot", "hammer", "shovel") as stimuli. The words "big" and "small" are paired with every noun, "ferocious" and "gentle" are paired with the animal nouns, "light" and "heavy" are paired with tools, and "rotten" and "tasty" are paired with food nouns. We name these adjective groups size, danger, manipulability and edibility. We also included the words "the" and the word "thing" to isolate the semantics of nouns and adjectives respectively. In total, there are 16 words, and 38 word pairs (phrases). Though some of the phrases start with a determiner (the), for simplicity, throughout this chapter we will refer to all 38 phrases as adjective-noun phrases. For a complete explanation of the experimental design, see Appendix A.

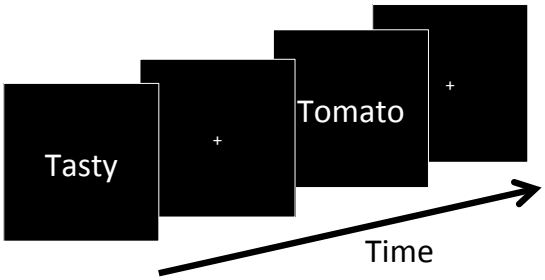


Figure 5.1: The paradigm used to collect MEG data to study adjective-noun phrases in the brain. Adjective-noun phrases were presented one word at a time with a fixation cross both between words and between phrases. The first word of the phrase appears at 0 seconds, and disappears at 0.5 seconds, the second word is visible 0.8s-1.3 seconds. The first words in successive phrases are 3 seconds apart.

Because of the experimental setup, there is a strong correlation between the adjectives and nouns in our data. For example, the word "rotten" only appears with food words "tomato" and "carrot". For this reason, we must be careful to avoid analyses that purport to decode a property of the adjective, but actually decode a correlated property of the noun (and vice versa).

To help subjects retrieve adjective-noun phrase semantics in a consistent way, subjects were shown the 38 phrases before the MEG recording session, and asked to plan what they would envision during the experiment: an exemplar (or exemplars) of the adjective noun phrase. While in the MEG scanner, the words were shown for 500 ms, with 300 ms of fixation between the words of a phrase, and 3 seconds total time between the onset of the first word in two consecutive phrases. The first word appears at 0 seconds, and disappears[1] at 0.5 seconds. The second word appears at 0.8s and disappears at 1.3 seconds. See Figure 5.1 for a pictorial representation of the paradigm. To ensure subjects were engaged during the experiment, 10% of the stimuli were adjective-adjective pairs (oddballs), for which the subjects were instructed to press a button with their left hand. Due to multiple word senses, the word "light" was not used in the adjective-adjective oddballs. Neither the adjective-adjective trials, nor the adjective-noun trial immediately following the oddball were used for analysis.

5.2.1 Data Acquisition and Preprocessing

All 9 subjects gave their written informed consent approved by the University of Pittsburgh (protocol PRO09030355) and Carnegie Mellon (protocol HS09-343) Institutional Review Boards. MEG data were recorded using an Elekta Neuromag device (Elekta Oy). The data were acquired at 1 kHz, high-pass filtered at 0.1 Hz and low-pass filtered at 330 Hz. Eye movements (horizontal and vertical eye movements as well as blinks) were monitored by recording the differential activity of muscles above, below, and beside the eyes. At the beginning of each session, the position of the subject's head is recorded with four head position indicator (HPI) coils placed on the subjects scalp. The HPI coils, along with three cardinal points (nasion, left and right pre-

[1] we will refer to the appearance and disappearance of stimuli as the onset and offset of stimuli.

auricular), were digitized into the system to allow for head translation to normalize data collected in different blocks.

The data were preprocessed using the Signal Space Separation method (SSS) (Taulu et al., 2004; Taulu and Simola, 2006) and temporal extension of SSS (tSSS) (Taulu and Hari, 2009) to remove artifacts and noise unrelated to brain activity. In addition, we used tSSS to realign the head position measured at the beginning of each block to a common location. The MEG signal was then low-pass filtered to 50 Hz to remove the contributions of line noise and down-sampled to 200 Hz. The Signal Space Projection method (SSP) (Uusitalo and Ilmoniemi, 1997) was used to remove signal contamination by eye blinks or movements, as well as MEG sensor malfunctions or other artifacts. Each MEG repetition starts at the onset of the first word of the phrase, and ends 3000 ms later, for a total of 3 seconds and 600 time points of data per sample. MEG recordings are known to drift with time, so we corrected the data by subtracting the mean signal amplitude during the 200ms before stimulus onset, for each sensor/repetition pair. Because the magnitude of the MEG signal is very small, we multiplied the signal by 10^{12} to avoid numerical precision problems. During subsequent behavioral tests (discussed in Section 5.6.1) it was found that phrases containing the noun "thing" were inconsistently judged by human subjects, and so the 8 phrases containing the noun "thing" were omitted from further analysis, leaving a total of 30 phrases for analysis.

After processing, the MEG data for each subject is composed of 20 repetitions for each of the 30 phrases. Each repetition has a 600 dimensional time series for each of the 306 sensors. For each subject, we average all 20 repetitions of a given phrase to create one data instance per phrase, 30 instances in all. The data set is now $30 \times 306 \times 600$.

MEG sensors are known to drift over time, so each of the phrases are baseline corrected. The average of the MEG signal in the 200ms before the onset of the first stimuli is subtracted from the full MEG time course on a per-sensor and per-phrase basis.

5.2.2 Source Localization

In order to transform MEG sensor recordings into estimates of neural activity localized to areas of the brain, we used a several step process. First Freesurfer (`http://surfer.nmr.mgh.harvard.edu`) was used to construct a 3D model of each subject's brain, based on a structural MRI. Freesurfer was used to segment the brain into ROIs based on the Desikan-Killiany Atlas. Then the Minimum Norm Estimate method (Hämäläinen and Ilmoniemi, 1994) was used to generate estimates from sources on the cortical sheet, spaced 5mm apart. Covariance of sensors was estimated using $\sim$ 2 minutes of MEG recordings collected without a subject in the room (empty room recordings) either directly before or after the subject's session. Source localization resulted in approximately 12000 sources per subject derived from the 306 MEG sensor signals.

5.3 Prediction Framework

To study adjective noun composition in the brain we have devised a simple prediction framework. Cross-validation is performed independently for each subject, wherein a subset of the 30 instances is withheld during training, and the withheld instances are used to test the framework's

predictions. This hold out and test procedure is repeated multiple times. For each of the decoding tasks described in Section 5.1.1, every phrase is associated with a vector of numbers which represent some semantic facet(s) of the phrase. The elements of this semantic vector become the targets (s) in the prediction framework. The semantic vector may be multi-dimensional or have only a single dimension. Depending on the task, each element of the semantic vector may be binary-valued, or contain a continuous value.

To predict each of the dimensions of the semantic vector, we trained an L_2 regularized (ridge) regressor $\hat{\beta}$:

$$\begin{aligned}\hat{\beta} &= \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} (s^{(i)} - \sum_{j=1}^{P} \beta_j x_j^{(i)})^2 + \lambda \sum_{j=1}^{P} \beta_j^2 \right\} \\ &= \underset{\beta}{\operatorname{argmin}} \left\{ ||S - X\beta||_F^2 + \lambda ||\beta||_F^2 \right\} \end{aligned} \tag{5.1}$$

where $s^{(i)}$ is the true value in one semantic dimension for training instance i, $x_j^{(i)}$ is one element of the matrix X, and represents the value for training instance i at a point j in MEG sensor time space, and λ is a regularization parameter that controls overfitting in β. We append a column of all 1 to X to allow us to incorporate a bias term. Each semantic dimension has its own $\hat{\beta}$. Initially, λ was tuned separately for each semantic dimension using leave one out cross validation (LOOCV) over the training data, but it was found to be very stable, so a set value of $\lambda = 10^{-6}$ was used to save computation time.

The regularized regressor in Equation 5.1 has a closed form solution:

$$\hat{\beta} = (XX^T + \lambda I)^{-1} X^T Y \tag{5.2}$$

MEG data was normalized during training so that each time-sensor feature has mean 0 and standard deviation 1. Let X_{train} be the subset of $X \in \mathbb{R}^{N \times p}$ used for training, where $p = 306 \times 600$ is the concatenation of all MEG sensor time series. During each fold of cross validation, the mean and standard deviation (μ and σ) are calculated for each column $j = \{1 \dots p\}$ of the training data X_{train} only. These factors are then used to normalize both train and test MEG data:

$$X(:, j) = \frac{X(:, j) - \mu}{\sigma} \tag{5.3}$$

where $X(:, j)$ represents one column of the full data matrix. Unless otherwise specified, MEG sensor data was used for all experiments.

5.3.1 The 2 vs. 2 Test

For each of the words in our study, we have m-dimensional semantic feature vectors $\mathbf{s_w} = \{s_1 \dots s_m\}$. The semantic vector assigned to the phrase may be either for the adjective, the noun, or the phrase, depending on the analysis being performed.

We train m independent functions $f_1(X) \rightarrow \hat{s}_1, \dots, f_m(X) \rightarrow \hat{s}_m$ where $\hat{s}_i$ represents the predicted value of a semantic feature. For these experiments, f is the regressor from Equation 5.1. We combine the output of $f_1 \dots f_m$ to create the final predicted semantic vector

$\hat{\mathbf{s}} = \{\hat{s}_1 \ldots \hat{s}_m\}$. We then define a function $d(\mathbf{s}, \hat{\mathbf{s}})$ that quantifies the dissimilarity between two semantic vectors. Any distance metric could be used here; we will use cosine distance:

$$\begin{aligned} d(\mathbf{a}, \mathbf{b}) =& 1 - cos(\mathbf{a}, \mathbf{b}) \\ =& 1 - \frac{(\mathbf{a} \cdot \mathbf{b})}{||\mathbf{a}|| \, ||\mathbf{b}||} \end{aligned}$$

Where $cos(\mathbf{a}, \mathbf{b})$ is the cosine of the angle between vectors $\mathbf{a}$ and $\mathbf{b}$.

To test performance we use the **2 vs. 2 test**. For each test we withhold two adjective-noun phrases and train regressors on the remaining 28. This is similar to leave one out cross-validation (LOOCV), but we leave out two words. We use the regressors f and MEG data from the held out phrases to predict two semantic vectors. The task is to choose the correct assignment of predicted vectors $\hat{\mathbf{s}}_\mathbf{i}$ and $\hat{\mathbf{s}}_\mathbf{j}$ to true semantic vectors $\mathbf{s}_\mathbf{i}$ and $\mathbf{s}_\mathbf{j}$. We will make this choice by comparing the sum of distances for the two assignments:

$$d(\mathbf{s_i}, \hat{\mathbf{s}}_\mathbf{i}) + d(\mathbf{s_j}, \hat{\mathbf{s}}_\mathbf{j}) \overset{?}{<} d(\mathbf{s_i}, \hat{\mathbf{s}}_\mathbf{j}) + d(\mathbf{s_j}, \hat{\mathbf{s}}_\mathbf{i}) \tag{5.4}$$

There are (30 choose 2) $= 435$ distinct 2 vs. 2 tests. Amongst those 435 tests, 51 share the same adjective and 60 share the same noun. **2 vs. 2 accuracy** is the number of 2 vs. 2 tests with correct assignments divided by the total number of 2 vs. 2 tests.

The 2 vs. 2 test is advantageous because it allows us to use two prediction per test, resulting in a higher signal-to-noise ratio; two weak predictions can still result in a correct assignment of true to predicted phrase vectors. Under the null hypothesis that MEG data and semantics are unrelated, two chance predictions will not increase the likelihood of a correct 2 vs. 2 assignment. Thus, the 2 vs. 2 test allows us to compute a better estimate of significantly above-chance performance, even in the face of noisy brain image data.

5.3.2 Classification Accuracy

Classification is used when we wish to differentiate between discrete groups of words (e.g. adjective attribute type classification). Much of the prediction framework is re-used for classification. The same regressor f is trained, but the output is a single value rather than a vector of values. For our classification tasks, the single value predicted by f is binary. For example, we will predict whether the first word of a phrase is the word "the" or an adjective. For this task, all phrases starting with "the" will be assigned the value 1, all others: 0.

Each of our classification tasks use binary (one-vs-all) labels; if we were to use the 2 vs. 2 test here, most of the 2 vs. 2 pairs would have identical labels, and so no discrimination could be made. Instead, we use LOOCV, and use a training set with 29 instead of 28 phrases. We also use an alternative performance metric, classification accuracy: the percentage of instances for which the correct class assignment was chosen.

5.3.3 Significance Testing

To determine statistical significance thresholds, we will use permutation and the Benjamini-Hochberg-Yekutieli procedure to control for multiple comparisons. The permutation test involves

shuffling the data labels (words) and running the identical prediction framework (cross validation, training β, predicting $\hat{s}$, computing 2 vs. 2 or classification accuracy) on the permuted data. When we do this many hundreds of times, we approximate the null distribution under which the data and labels have no decodable relationship. From this null distribution we calculate a p-value for performance when training on the correct assignment of words to MEG data. In the experiments that follow we will train and test multiple predictors on multiple time windows. To account for the multiple comparisons performed over the time windows we will use a variant of False Discovery Rate (FDR) correction: The Benjamini-Hochberg-Yekutieli procedure. This FDR procedure requires sorting the p-values obtained for a particular experiment (e.g. the 2 vs. 2 accuracy over many time windows). The FDR threshold is the smallest p-value to satisfy the constraint:

$$p_i > \alpha \frac{i}{m \sum_{j=1}^{m} \frac{1}{j}} \tag{5.5}$$

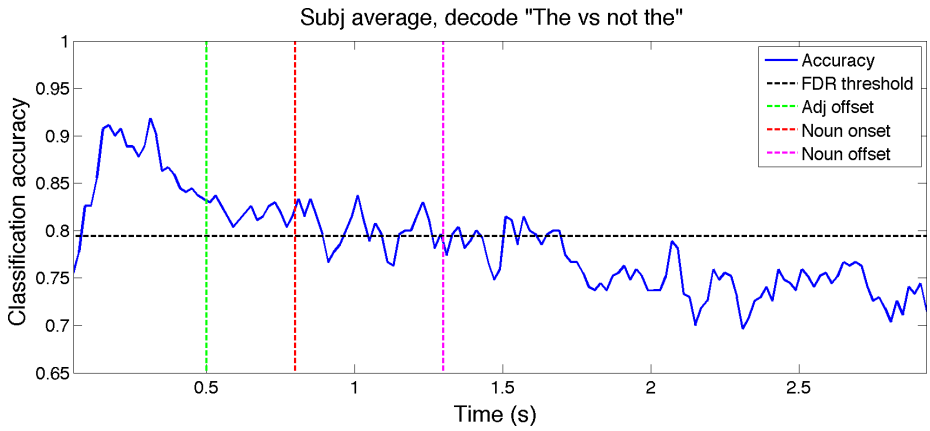
where i is the index of a sorted p-value, m is the total number of p-values to be tested, alpha is the significance threshold (as is typical, we use $\alpha = 0.05$), and $\sum_{j=1}^{m} \frac{1}{j}$ is a correction for arbitrary correlation amongst the tests. FDR correction allows us to separate out the long tail of statistically significant tests from tests drawn from the null. The first FDR test (p_1) is equivalent to Bonferroni correction. If the first test passes, each subsequent test becomes slightly more permissive; the FDR threshold adapts to the data being tested.

Permutation tests are incredibly computationally expensive, as they must be run hundreds of times to build a good estimate of the null distribution. We can speed up our permutation tests by noting that the solution to the regression problem (Equation 5.1) is a product of a function of the data ($(XX^T + \lambda I)^{-1}X^T$) and a function of the labels (Y). Thus, when computing permutation test results, we need only compute the function of the data one time, and can reuse the solution to compute many permutation test weight matrices.

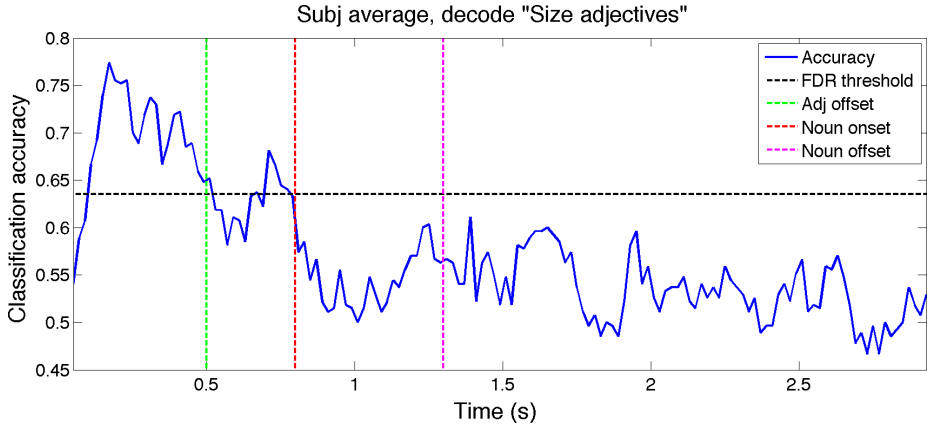
5.4 Decoding the Adjective Attribute Type

We consider decoding the attribute type of the adjective being processed. For example, can we tell if the adjective being processed is size-related (big, small) or edibility-related (tasty, rotten)? For this experiment, we train 5 independent regressors corresponding to each of the 4 adjective categories (edibility, manipulability, danger and size), and one for the determiner "the". Each phrase has a non-zero value for exactly one of the classification tasks. We train regressors with the LOOCV framework for classification described above.

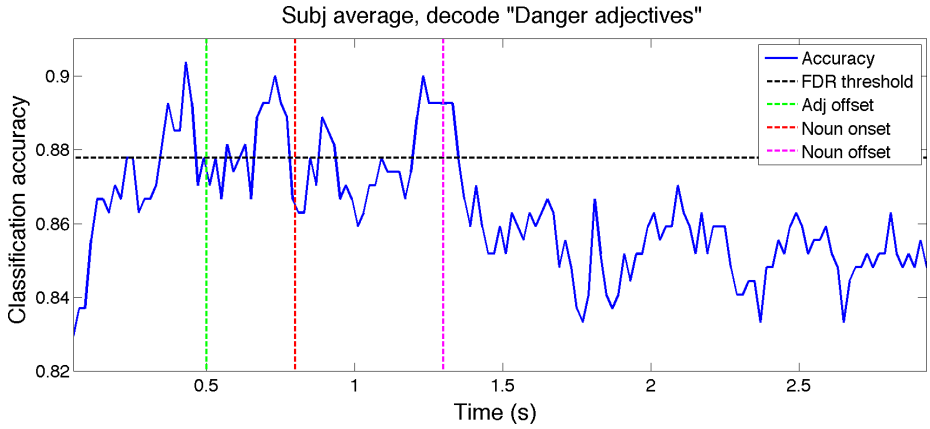
For three of the adjective type classes (edibility, manipulability and danger) only $4/30$ phrases fall into the positive category, and so a naïve predictor that always predicts the majority class will perform very well, correctly assigning values 87% of the time. In contrast, the majority class for "the" represents 80% of the training phrases, and 60% for "size", making it slightly easier to outperform the naïve predictor for these classification tasks. The mean of the null distribution for each of these classification tasks is a few percent below the naïve majority class predictor: 75% vs 80% for "the", 53% vs 60% for "size", and 85% vs 87% for each of "edibility",



(a) Classification accuracy, as a function of time, for decoding if the first word of a phrase is “the”.



(b) Classification accuracy, as a function of time, for decoding if the first word of a phrase is a size-related adjective.



(c) Classification accuracy, as a function of time, for decoding if the first word of a phrase is a danger adjective.

Figure 5.2: Classification accuracy, averaged over all 9 subjects, as a function of time for the task of decoding the adjective attribute type from MEG signal. Time windows are 100ms wide and overlap by 80ms with adjacent windows. The value on the x axis represents the center of the time window. Time 0 is the onset of the adjective, green lines indicates the offset of the adjective, red: onset of noun, magenta: offset of noun.

"manipulability" and "danger". This illustrates the known pessimistic nature of a cross-validated estimate of accuracy.

Note that when we train a one-vs-all classifier for either "edibility", "manipulability" or "danger" using LOOCV, we could still produce results that rely on the confound that the noun is highly correlated to the adjective. For example, each edibility adjective appears only with food nouns. However, each food noun also appears with both size adjectives and "the", so using the adjective-correlated nouns to make predictions about the adjective-type will result in 3 incorrect predictions for every 2 correct, a losing strategy.

Because of the excellent time resolution, we can use MEG data to examine the time course of decodability for each of the decoding tasks. For this, we train and test on 100ms windows of MEG signal, across all sensors. The window is shifted by 20 ms and we train and test again until the end of the window reaches 2.98 s post adjective onset. This allows us to plot decodability as a function of time. In all of the graphs in this chapter, green, red and magenta lines signify the offset of the adjective, onset of noun and offset of noun, respectively. Black dashed lines are FDR thresholds.

Figure 5.2 shows the decodability of adjective type over time. Amongst "edibility", "manipulability" and "danger", only "danger" and "manipulability" produce classifications significantly above chance at any point in time. Manipulability produces only one time point significantly above chance for a window centered at 0.17s after adjective onset, so only the plot for danger adjectives is pictured in Figure 5.2. Danger's representation has fairly sustained decodability, remaining decodable until ~ 1.4 s, just after the offset of the noun. Discerning adjective-noun phrases from phrases that begin with "the" has peak accuracy at around 0.3 s after the onset of the adjective, and sustained decodability until ~ 1.65 s. Size-related adjectives are decodable quite early, and the decodability drops below the FDR threshold as soon as the noun stimuli is perceived.

5.5 Decoding Adjective and Noun Semantics

Table 5.1: 2 vs. 2 accuracy for decoding the adjective or noun during the time the adjective or noun is being presented. When decoding the adjective, only 2 vs. 2 pairs where the noun is shared are considered. When decoding the noun, only 2 vs. 2 pairs where the adjective is shared are considered. Note that predicting the noun during the time the adjective is presented gives chance performance ($\sim 50\%$), as the noun has not yet been perceived by the subject.

Task	Adjective presentation (0-0.8s)	Noun presentation (0.8-1.6s)
Decode Adjective	93.33	76.11
Decode Noun	48.80	89.54

For each of the words in our study, we have semantic feature vectors $\mathbf{s} = \{s_1 \dots s_m\}$. The semantic vectors are from Fyshe et al. (2013)[2], and are based on the dependency relationships

[2]Vectors available for download from `http://www.cs.cmu.edu/~afyshe/papers/conll2013/`

incident on the word of interest, as calculated over many sentences (see Section 3.3 for more details). We use the first $m = 500$ SVD dimensions to summarize the dependency statistics. We will use the 2 vs. 2 test to explore the properties of adjective and noun semantics in the brain.

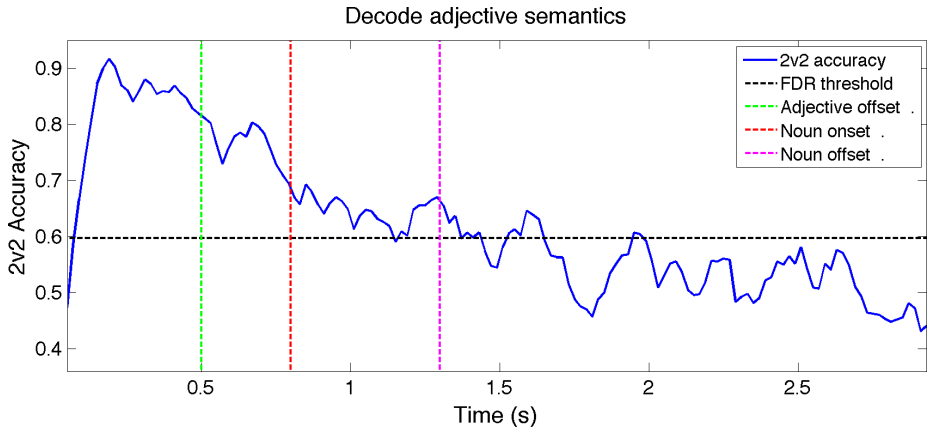
Recall the correlation between adjectives and nouns in our experimental paradigm. To avoid reporting results that rely on this confound, we only consider decoding the adjective or noun when the other word (noun or adjective, respectively) is shared in the 2 vs. 2 pair. That is, when we encounter a 2 vs. 2 pair that contrasts adjectives "rotten" and "big", we will do so only when the noun is the same for both phrases (e.g. "tomato"). If our prediction framework leverages the correlated semantics of the noun to decode the adjective, it would be of no use for differentiating between these phrases, as the noun (and thus the noun semantics) are identical. When decoding the noun, the adjective in the 2 vs. 2 pair is shared, so 2 vs. 2 accuracy should not be significantly above chance before the onset of the noun, as the semantics of the adjective will not help to differentiate the 2 vs. 2 pair. Thus we can be sure that we are not relying on any correlations between adjectives and nouns for the analyses in this section.

Previous work has shown that one can decode the semantics of the noun a person is reading based on the MEG signal recorded during reading (Sudre et al., 2012; Chan et al., 2011). Table 5.1 shows that we too can decode each word of the phrase during the time it is presented, even though our words appear in phrases. In addition, Table 5.1 shows that we can decode the adjective being presented during the time the noun is being read. Thus, some semantic representation of the adjective persists even while the noun is being read. Note that we should not be able to predict the noun during the time the adjective is presented, as the noun has not yet been perceived by the subject. As expected, noun decoding during adjective presentation is close to chance performance (50%).

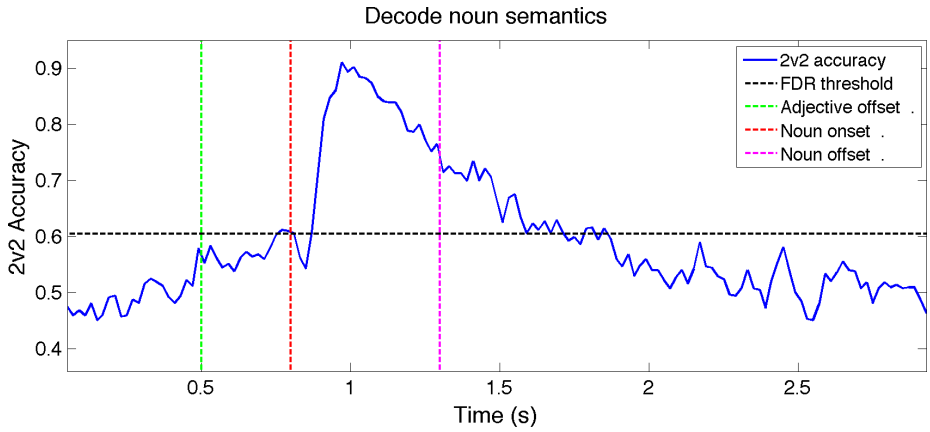
Results for decoding the adjective and noun as a function of time in time appear in Figure 5.3, and are FDR corrected for multiple comparisons. Here we can see that the adjective remains decodable until well after the onset of the noun, dipping below the FDR threshold for the first time at about 1.2s after the onset of the adjective (0.4 s after the onset of the noun). Adjective decodability is significantly above chance at several more time points after the first below-FDR dip, and is significantly above chance for the last time around 2s after the onset of the adjective. Thus, we can decode the adjective during the adjective and noun presentation, and also for a prolonged period after the noun stimuli has disappeared. This implies that there is a neural encoding associated with the adjective that is sustained during the entire phrase presentation, as well as after the phrase has been presented (phrase wrap-up period). Note that from these results we cannot tell if the encoding of the adjective changes over time, we can only infer that there exists a reliable encoding during each significantly above-chance time window. After its initial peak around 0.95s, the decodability of noun semantics dips below the FDR threshold for the first time at ~ 1.7s, and is no longer significantly above chance after ~ 1.8s.

5.5.1 Consistency of the Neural Code in Time

How consistent in time is the sustained neural code of a word? For example, does the neural code for the adjective during adjective presentation resemble the neural code used to encode the adjective during noun presentation, or during the phrase wrap-up period? To test this we create what we call a train-test time matrix (TTM), which uses the prediction framework described in



(a) 2 vs. 2 accuracy for decoding the semantics of the adjective, as a function of time. To avoid confounds, only 2 vs. 2 pairs where the noun is shared are used to calculate accuracy.



(b) 2 vs. 2 accuracy for decoding the semantics of the noun, as a function of time. To avoid confounds, only 2 vs. 2 pairs where the adjective is shared are used to calculate accuracy.

Figure 5.3: 2 vs. 2 decoding as a function of time for the task of decoding the adjective or noun from MEG signal, based on its corpus-derived semantic vector. Time windows are 100ms wide and overlap by 80ms with adjacent windows. Time 0 is the onset of the adjective, green lines indicate the offset of the adjective, red: onset of noun, magenta: offset of noun.

Section 5.3, but mixes training and test data from different time windows. The entry at (i, j) of a TTM ($T_{i,j}$) contains the 2 vs. 2 accuracy when we train using MEG data from a time window centered at time i, and test using MEG data from a time window centered at time j. Thus, depending on the value of i and j we may be training and testing during time periods when the subject is viewing a different word type, or no visual stimuli at all. An example of one subject's time matrix for decoding adjective semantics can be seen in Figure 5.4. Along the y axis is the time used to train weight maps to predict the semantic vector, and along the x axis is the time used to test the weight map. For each TTM coordinate, the 2 vs. 2 test is performed: two words from the dataset are omitted from training and a different segment of the MEG signal is used to test. Data is normalized based on the feature-specific μ and σ of the training data. The diagonal of the TTM corresponds to training and testing with the same time window. For all TTMs, windows are 100ms wide and overlap by 80 ms with adjacent windows. The diagonal of a TTM averaged over all subjects is exactly to the result plotted in Figure 5.3.

TTMs for predicting the semantics of an adjective or noun, averaged over all subjects, appear in Figure 5.5 and 5.6. The color in each cell represents the $-ln(\text{p-value})$ for that particular 2 vs. 2 result, averaged over subjects. Dark blue TTM coordinates fall below the FDR threshold. To increase statistical power, off diagonal elements are combined into one significance test, as training on time i and testing on time j ought to yield similar results to training on time j and testing on time i. Thus, only tests above the diagonal line of the matrix are shown. The diagonal (where train and test time are identical) is omitted from this analysis. Figure 5.5a shows a clear pattern of significantly above chance decoding when we train during the time the adjective is presented and test late in time, or vice versa. Note that the striped pattern of these significantly above chance patches matches the strong diagonal lines we see in the TTM for subject D (Figure 5.4).

Figure 5.5b is identical to Figure 5.5a, but with thresholding to show train-time coordinates where adjective decoding is significantly below chance. Note that the significantly below chance periods also show an oscillatory pattern, and that the strongest group is clustered during the portion of the TTM corresponding to training on time during the noun presentation and testing on the adjective presentation, or vice versa.

Figure 5.6 shows the TTM for decoding noun semantics, averaged over subjects and FDR thresholded. Note the absence of a large off-diagonal pattern like that seen for decoding adjective semantics (Figure 5.5a). Also note that the diagonal band during noun presentation is much narrower than the diagonal band during adjective presentation in Figure 5.5a. In addition, there are very few significantly below chance TTM coordinates for decoding noun semantics (Figure 5.6b).

Now we investigate the brain areas that are responsible for the pattern we see in the TTMs for adjective semantics (Figure 5.5). For this exploration, each subject's source localized MEG data was divided into 6 gross ROIs (regions of interest) based on the Freesurfer parcellation. They are:

temporal temporal pole, transverse temporal, middle temporal, inferior temporal, superior temporal, bank of superior temporal sulcus, fusiform

inferior frontal gyrus (IFG) pars opercularis, pars orbitalis, pars triangularis

parietal inferior parietal, superior parietal, postcentral, supramarginal

occipital lateral occipital

frontal frontal pole, lateral orbitofrontal, medial orbitofrontal, rostral middle frontal, superior frontal, caudal middle frontal, precentral

limbic caudal anterior cingulate, isthmus cingulate, posterior cingulate, rostral anterior cingulate, parahippocampal

Because language is left-lateralized, we consider the left hemisphere only. Within each of these super-ROIs, we perform the same 2 vs. 2 train-test setup with FDR thresholding as in Figure 5.5, but use only the subset of the source localized MEG data associated with that ROI. The same statistical testing procedure is used here, corrected for the fact we are performing tests over 6 different ROIs.

ROI results appear in Figure 5.7. Amongst the 6 ROIs, we see that occipital, limbic and parietal regions contribute the most to off-diagonal accuracy. The off diagonal accuracy for the occipital ROI is particularly striking. Note that temporal and IFG make only small contributions to the off-diagonal accuracy; these regions have been implicated in previous studies of composition (Bemis and Pylkkänen, 2011, 2013b).

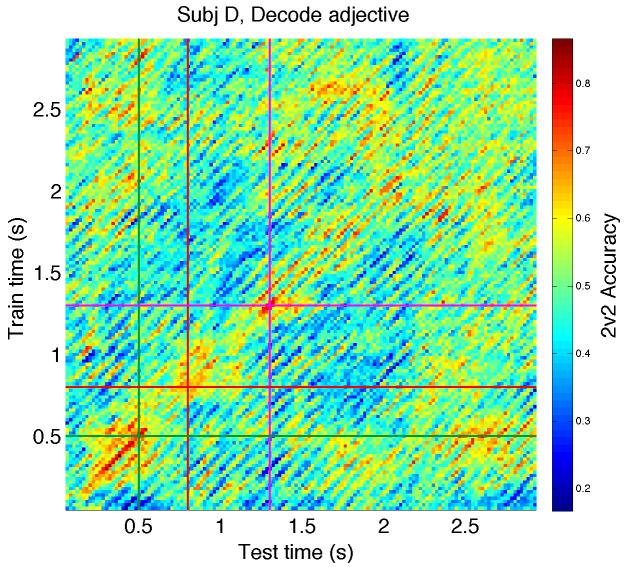
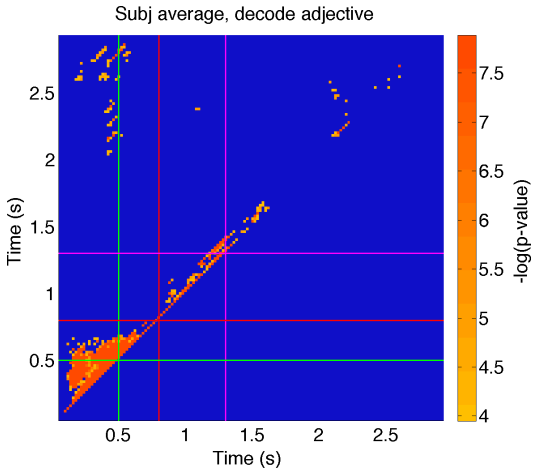


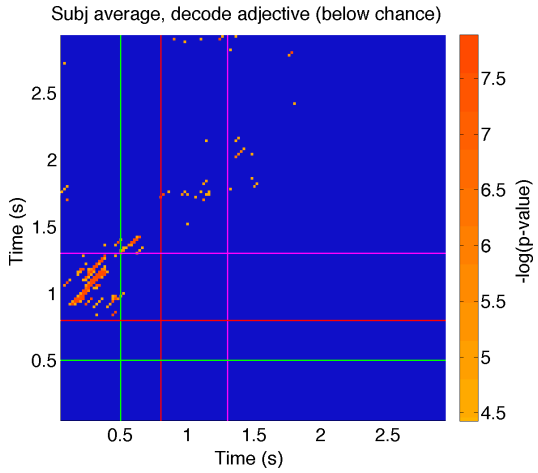
Figure 5.4: A Train Test Time Matrix for decoding adjective semantics for one subject (D). Green line is the offset of the adjective, red: onset of noun, magenta: offset of noun.

5.6 Decoding Phrasal Semantics

Now we turn to the task of decoding the semantics of the phrase, which represents elements of adjective and noun semantics, as well as their interactions. In Chapter 3 we described a system for learning a semantic representations for adjectives and nouns that allows for better approximation of phrasal meaning. One of the negative aspects of these compositional models is that the estimated phrasal representation is often very close to either the adjective or noun. In Section 3.4, we note that this effect is so strong that Turney (2012) included a special case in the dual space comparison function so that adjective and noun vectors would effectively not be

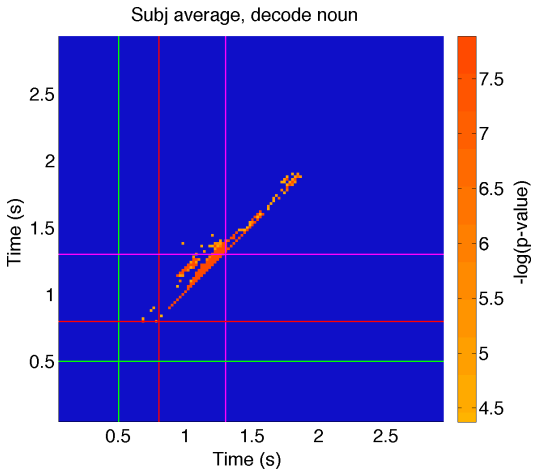


(a) FDR thresholded TTM showing significantly above chance time coordinates for decoding adjective semantics.

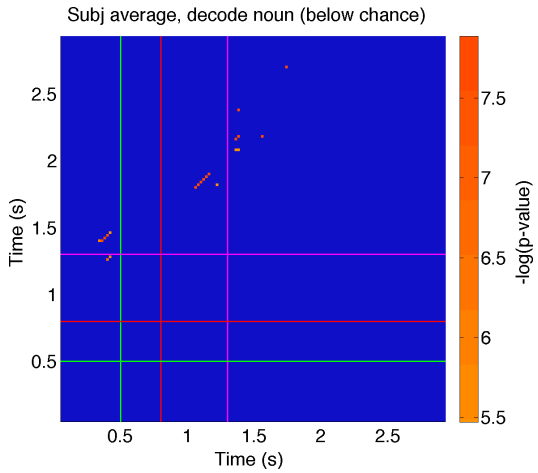


(b) FDR thresholded TTM showing significantly below chance time coordinates for decoding adjective semantics.

Figure 5.5: FDR thresholded TTMs for decoding adjective semantics from the MEG signal. Only above-diagonal coordinates are shown, as our FDR tests combine TTM coordinates i, j with j, i. Blue TTM coordinates are below the FDR threshold. Time windows are 100ms wide and overlap by 80ms with adjacent windows. Time 0 is the onset of the adjective, green lines indicates the offset of the adjective, red: onset of noun, magenta: offset of noun.



(a) FDR thresholded TTM showing significantly above chance time coordinates for decoding noun semantics.



(b) FDR thresholded TTM showing significantly below chance time coordinates for decoding noun semantics.

Figure 5.6: FDR thresholded TTM for decoding noun semantics from the MEG signal. Time windows are 100ms wide and overlap by 80ms. Only above-diagonal coordinates are shown, as our FDR tests combine TTM coordinates i, j with j, i. Blue TTM coordinates are below the FDR threshold. Time 0 is the onset of the adjective, green lines indicates the offset of the adjective, red: onset of noun, magenta: offset of noun.

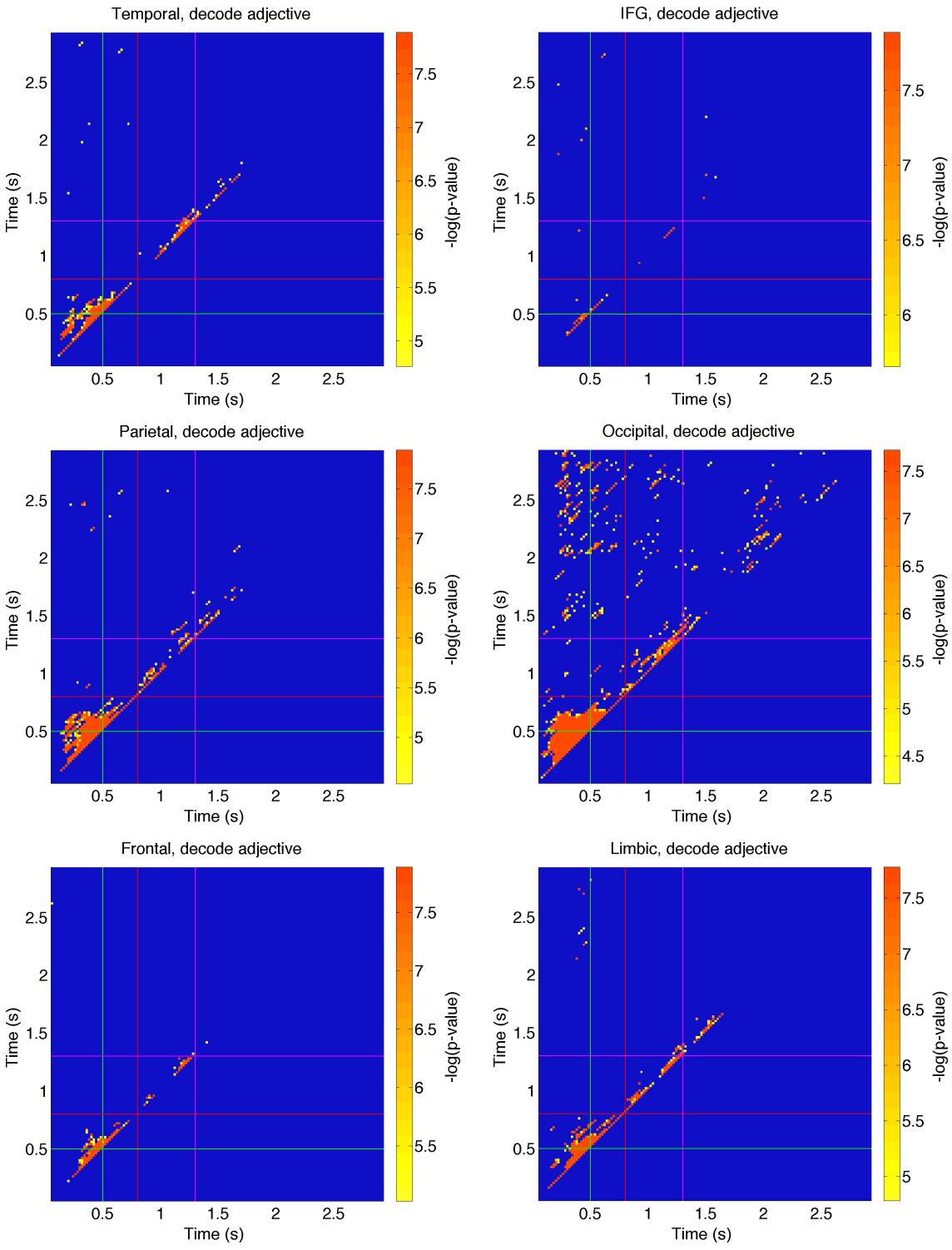


Figure 5.7: FDR thresholded TTMs for decoding adjective semantics using source localized data from 6 ROIs (temporal, inferior frontal gyrus (IFG), parietal, occipital, frontal and limbic). Only above-diagonal coordinates are shown, as our FDR tests combine TTM coordinates i, j with j, i. Blue TTM coordinates are below the FDR threshold. Time windows are 100ms wide and overlap by 80ms with adjacent windows. Time 0 is the onset of the adjective, green lines indicates the offset of the adjective, red: onset of noun, magenta: offset of noun.

considered during ranking. Because composed phrasal semantic vectors are so strongly tied to the single word units, we cannot justify using them for phrase decoding from brain data. To do so would make it impossible to prove that the algorithm is not just decoding the meaning of the noun or the meaning of the adjective.

5.6.1 Behavioral Data

To avoid the problem of phrasal vectors being too close to adjective or noun vectors we instead approximate phrasal meaning by collecting behavioral data on the phrases themselves. This also bypasses the issue that we do not actually know the function that combines adjectives and nouns to make phrasal semantics. We instead ask people to employ their own mental composition function and describe the output with a series of rating questions.

We asked four sets of questions on Mechanical Turk. Three were simple rating questions:

1. How likely is it that a person would eat a <noun phrase>?
2. How easy is it to pick up a <noun phrase>?
3. Is a <noun phrase> scary or friendly?

where the mechanical turk user was asked to answer the question on a rating scale $[1 \ldots 5]$. For example, 1 corresponds to "Very friendly" and 5 to "Very scary" for the third question above. Each question was answered by 5 people, and the median of the five answers was used as the final behavioral rating. The median ratings for the 30 phrases can be seen in Table 5.2.

The fourth set of questions are pairwise comparisons to contrast the physical size of the items referred to by the 38 phrases. For every pair of phrases we asked:

What is the size relationship between a <noun phrase 1> and a <noun phrase 2>?

The answers are ratings, with 1 corresponding to "A <noun phrase 1> is much smaller than a <noun phrase 2>", and 5 to "A <noun phrase 1> is much larger than a <noun phrase 2>". We did not assume that size ratings would be symmetrical, and collected pairwise ratings for both orderings of phrase pairs. We assumed that the size rating of each noun with itself would be the neutral rating (3). Each question was answered by 5 users, which resulted in $38 * 5$ answers per phrase, many of which contain redundant information. To assess the accuracy of these ratings, we calculated the mean answer, and applied Singular Value Decomposition to the resulting 38×38 matrix. The first SVD dimension produces a fairly good ordering of the phrases by size, which can be seen in Figure 5.8.

We included a section at the end of each question set for users to leave comments. The comments implied that answering questions about "adjective thing" phrases was difficult and unnatural, so we discarded these questions and answers from our dataset, as well as the corresponding MEG data.

If we combine all behavioral data for all phrases, we have a matrix that has 38 rows (one per phrase) and 41 columns (38 pairwise size ratings and 3 other adjective rating questions). We took this full behavioral matrix and applied SVD to it, retaining the dimensions such that the sum of their eigenvalues was responsible for $> 90\%$ of the total sum of eigenvalues, 12 dimensions in all. We use these vectors to perform the 2 vs. 2 test to differentiate phrases. To avoid confounds with the noun, we will consider only 2 vs. 2 pairs that share a noun.

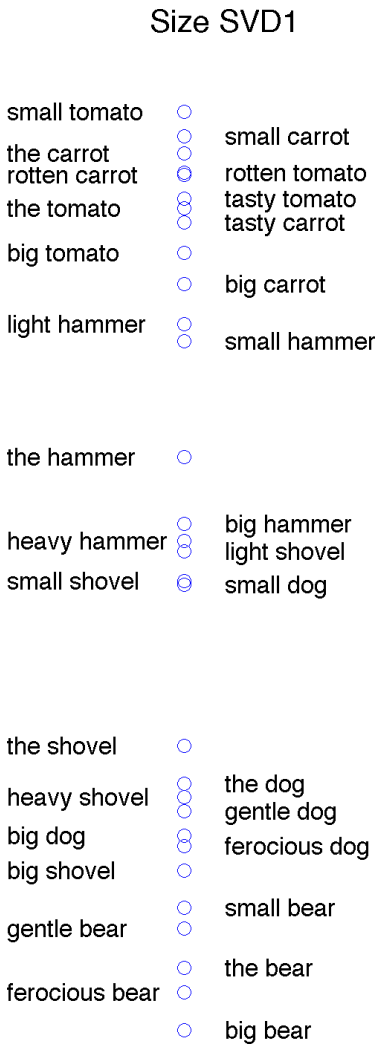
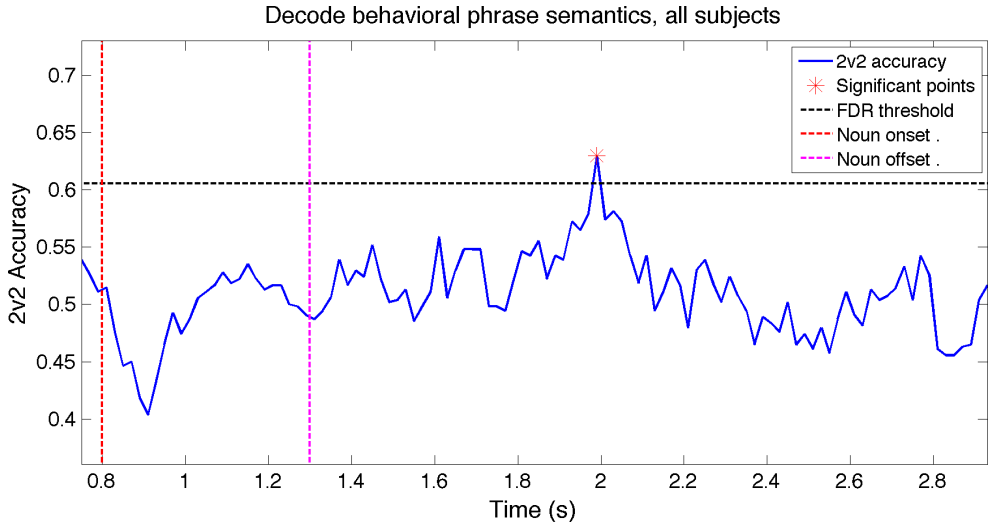


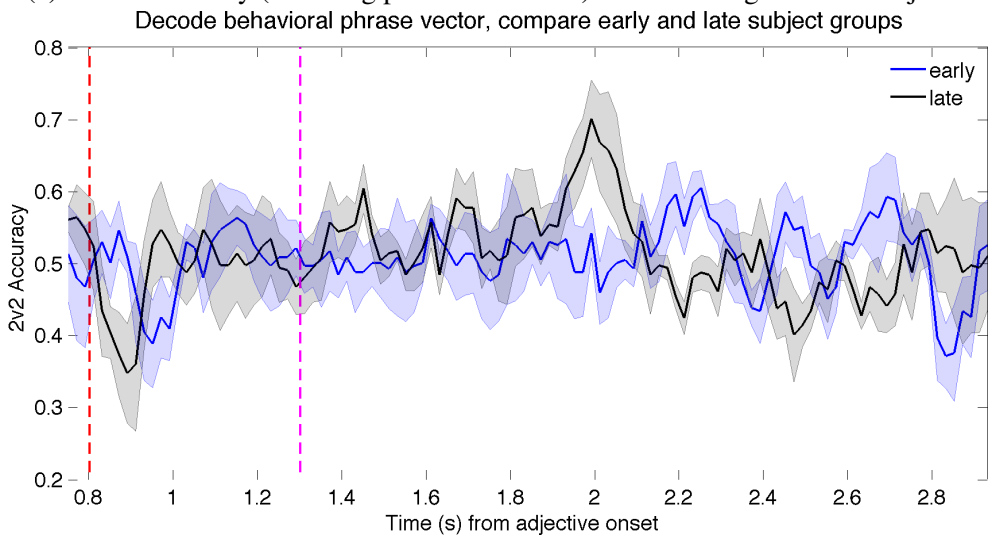
Figure 5.8: The 30 phrases used in this study, ordered by the first SVD dimension summarizing the behavioral size rating scores. Note that smaller objects appear at the top and larger towards the bottom.

Table 5.2: Behavioral rating scores for three question sets and the 30 adjective noun phrases (median over 5 mechanical turk users' responses). Ratings were on a scale $[1 \ldots 5]$.

Rating	Phrases
Edibility	
1	the dog, the bear, big dog, small dog, ferocious dog, gentle dog, big bear, small bear, gentle bear, rotten tomato, rotten carrot, the hammer, the shovel, big hammer, small hammer, heavy hammer, light hammer, big shovel, small shovel, heavy shovel, light shovel
2	ferocious bear
5	the tomato, the carrot, big tomato, small tomato, tasty tomato, big carrot, small carrot, tasty carrot
Manipulability	
1	the bear, ferocious dog, big bear, ferocious bear, gentle bear
2	big dog, small bear, big hammer, heavy shovel
3	the dog, heavy hammer, big shovel
4	gentle dog, the hammer, the shovel
5	small dog, the tomato, the carrot, big tomato, small tomato, tasty tomato, rotten tomato, big carrot, small carrot, tasty carrot, rotten carrot, small hammer, light hammer, small shovel, light shovel
Danger	
1	gentle dog
2	the dog, small dog, gentle bear
3	small bear, the tomato, the carrot, big tomato, small tomato, tasty tomato, rotten tomato, big carrot, small carrot, tasty carrot, the shovel, big hammer, small hammer, heavy hammer, light hammer, big shovel, small shovel, heavy shovel, light shovel
4	big dog, rotten carrot, the hammer
5	the bear, ferocious dog, big bear, ferocious bear

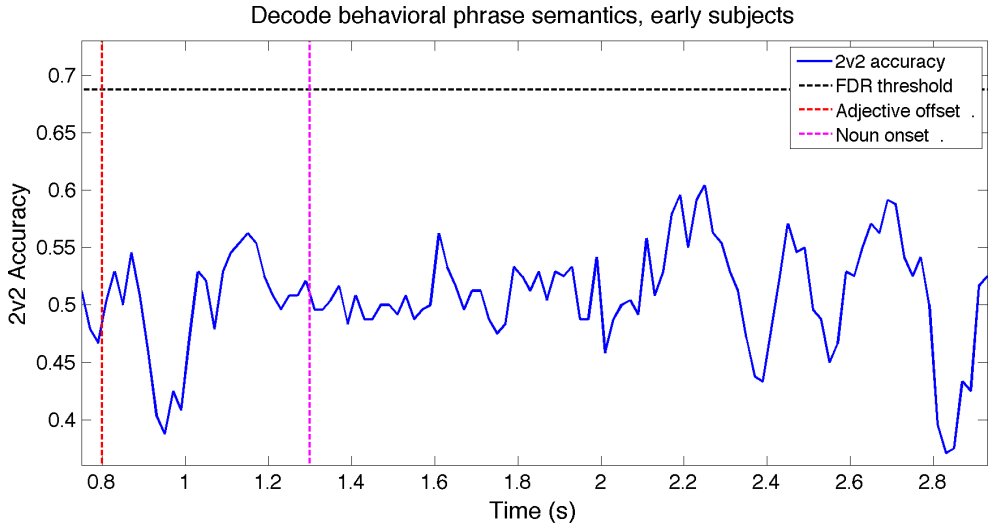


(a) 2 vs. 2 accuracy (decoding phrase semantics) for the average of all 9 subjects.

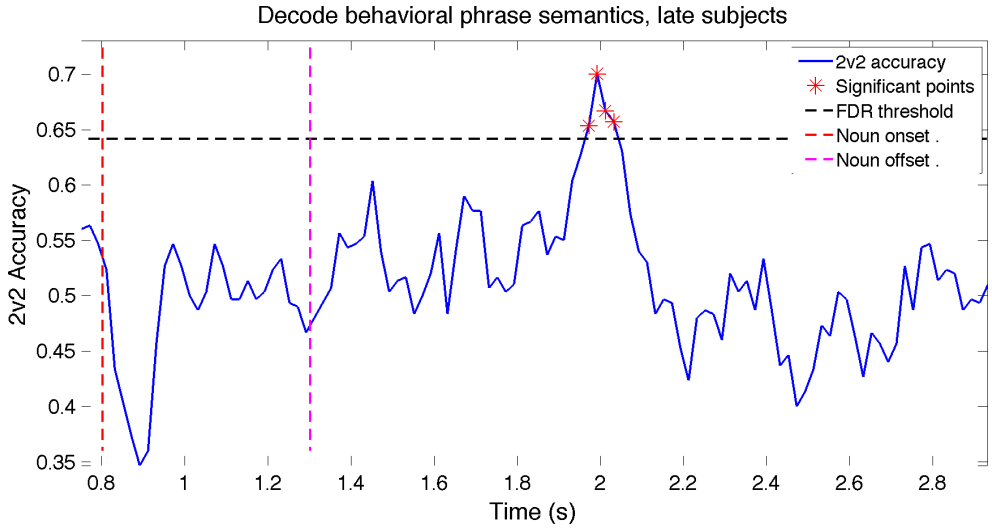


(b) 2 vs. 2 accuracy (decoding phrase semantics) for two subject groups created based on the timing of peak decodability of adjective semantics in off-diagonal coordinates of the TTM. Shaded area represents the standard error of the mean.

Figure 5.9: 2 vs. 2 accuracy for decoding phrase semantics. (a) average over all subjects: one significant point at 2s. (b) Decodability when subjects are divided into two groups based on the timing of their peak off-diagonal adjective decoding accuracy.



(a) 2 vs. 2 accuracy (decoding phrase semantics) for the average of 3 subjects that display early off-diagonal adjective accuracy.



(b) 2 vs. 2 accuracy (decoding phrase semantics) for the average of all 6 subjects that display late off-diagonal adjective accuracy.

Figure 5.10: 2 vs. 2 accuracy for decoding the phrase vector for groups defined by the peak off-diagonal adjective decoding accuracy in their TTM . The early accuracy group has no significant points; the late group has 4. There are 4 subjects in the early group, which leads to a slightly higher variance permutation test, and higher FDR threshold.

2 vs. 2 accuracy as a function of time for predicting the behavioral phrase vector appears in Figure 5.9a. There is one small peak above the FDR threshold at 2s post adjective onset. This aligns with the last significantly above chance peak in Adjective decodability in Figure 5.3a.

5.6.2 Subject-Dependent Variation in Phrase Decodability

We noticed that the TTMs for adjective semantics had some subject-dependent variability, especially in the timing of peak off-diagonal adjective decodability. Some subject's TTMs had highest 2 vs. 2 accuracy if trained on a window during adjective presentation and tested on a window 1.3-2s. Some subjects had higher accuracy if we tested on a window after 2s. We hypothesized that this inter-subject variability might affect our ability to decode the phrase. We separated the subject pool into two groups, early (off-diagonal adjective peak $<$ 2s, mean 1.72 ± 0.28) and late (off-diagonal adjective peak $>=$ 2s, mean 2.69 ± 0.08). Phrase decodability as a function of time for the two groups appears in Figure 5.9b, clearly showing differences in peak phrase decodability. Though the early group has off-diagonal peak adjective accuracy around 1.7s, we actually see peak *phrase* decodability around 2.2s for these subjects, during which time the decodability for late subjects is significantly lower.

Figure 5.10 shows early and late phrase decodability with FDR thresholds. The late group produces 4 points significantly above chance, centered at 2s; the early group has no significant points. This is partially due to differences in group size: the early group has 4 members, the late group has 5. The permuted 2 vs. 2 accuracy averaged over 4 subjects will have slightly higher variance than one averaged over 5 subjects, which will result in a higher FDR threshold. The early group also has higher inter-group variance, as evidenced by the higher variance amongst the subject-specific off-diagonal adjective TTM peak. This is not conclusive evidence that inter-subject variability has an effect on phrase decodability, as we used adjective decodability (which is probably correlated with phrase decodability) to divide the subjects. But, it is reasonable to assume that there will be more inter-subject variability as we begin to explore more complex cognitive tasks, like semantic composition.

5.7 Discussion

Here we compare and contrast the results from Section 5.4-5.6 to build a picture of how the brain is represents and processes adjective noun semantics.

5.7.1 Timing of Decodability

Adjective attribute decodability has several interesting properties. Note that all phrases in this experiment that do not start with "the" start with an adjective. Thus, when we decode if the phrase begins with the word "the" (Figure 5.2a), we are identifying the *lack* of adjectival processing. "The" decodability is sustained until well after the offset of the noun ($\sim$ 1.6s). This is very similar to the decodability of adjective semantics (Figure 5.3a), which is mostly above the FDR threshold until about the same time, and is also similar to the diagonal strip in the TTM for adjective decodability in Figure 5.5a.

Size adjectives have the interesting property that their impact on the meaning of a noun is strongly affected by the noun itself. For example a big tomato is still smaller than a small bear. In addition, the absolute size difference between a small tomato and big tomato is smaller than the absolute size difference between a small and large bear. Thus, size adjectives cannot be fully processed until the noun's meaning is encountered, at which point the magnitude of the change is calculated based on the noun. In light of these facts, note that size attribute decodability (Figure 5.2b) drops below the FDR threshold as soon as the noun is presented. This implies that there is a neural representation for the intent to modify the size of an object (in either the positive or negative direction) that is maintained until the noun is encountered, but that neural representation vanishes once the noun appears, at which time the size adjective can be fully processed and applied to the noun.

Adjectives that modify the dangerousness of a noun are the only noun-specific adjective type that have more than one significantly above chance time point. (See Figure 5.2c). The fact that danger-attribute adjectives are decodable significantly above chance, despite the minimal number of examples is intriguing. There is likely an evolutionary advantage to being able to quickly and reliably process the threatening nature of an animal. Perhaps this is the reason that danger is decodable, whereas other adjective types are less decodable.

5.7.2 Adjective Semantics in Early and Late Time Windows

Recall that we can decode the adjective well after the the noun stimuli has left the screen, and that there is a neural encoding of the adjective late in the MEG trial that is consistent with the representation recalled when the adjective is first perceived.

Figure 5.5a (adjective semantics TTM), shows significantly above chance decodability for coordinates close to the diagonal until about 1.7s after the onset of the adjective, and then sparsely between 2-2.5s. The diagonal band is widest during the time the adjective stimulus is being viewed. Another striking pattern is the large off-diagonal patches of significant coordinates that appear when we train on a time point after 2s (when no words are on the screen) and test on a time point during adjective presentation or intra-phrase fixation (0.2-0.8s). This pattern implies that the neural code that appears while the adjective is being read and understood is similar enough to the pattern that appears during the phrase wrap-up period that the data from the two time periods can be exchanged during train and test and still perform with significantly above chance accuracy. This is also evidence that the small peak of significantly above chance adjective decoding at 2s in Figure 5.3a is a real effect, as it corresponds to a point in time where the adjective encoding is consistent with the encoding during the adjective's initial perception and understanding. This effect could be due to the intersective nature of the majority of our adjectives. The meaning of the edibility, manipulability and danger adjectives is largely unaffected by the semantics of the nouns they are paired with in this experiment. Thus, the composed phrasal meaning may use a neural encoding that is very similar to that of the adjective.

Note the oscillatory nature of the off-diagonal patch in Figure 5.5a: the patches run in diagonal lines parallel the main diagonal of the TTM . Note also the highly oscillatory nature of the decodability results in the adjective semantics TTM for a single subject (Figure 5.4). We will return to this pattern in Section 5.7.5.

Though not shown, high accuracy off-diagonal patches of the TTM tend to be stronger when

we train on later time points and test on earlier time points. It has been shown that L2 regularization will spread learned weights amongst correlated features in a feature vector (Zou and Hastie, 2005). Thus, under L2 regularization, training on a low noise signal will cause weights to be evenly distributed amongst a set of correlated variables. But, in a higher-noise scenario, the set of correlated variables may appear smaller due to noise, and thus the weights would be spread amongst a smaller set of features. If the neural encoding in later time periods is more noisy, L2-regularization could produce a weight map more suited for the possibly lower-noise neural encoding that appears at earlier time points. If the weight map is instead trained on low-noise and tested on high-noise data, features that will suffer from noise at later time points cannot be identified and down-weighted in the weight map, leading to poorer performance. This may also explain why there are no points in Figure 5.3a above the FDR threshold after 2s. Perhaps the signal at late time points is noisy enough that it is better to test or train with data from a lower noise time period than from the same (high noise) time period. This can be confirmed in the single subject adjective semantics TTM (Figure 5.4), which has several high 2 vs. 2 accuracy stripes in the upper left and lower right corners, but very few on the diagonal in the upper right quadrant. This could imply that points after 2s in the time course of adjective semantics decoding (Figure 5.3a) are actually significant, but our FDR criteria is too conservative, or that more subjects are required to see the effect.

The off-diagonal patch in Figure 5.5a shows that the code for the adjective appears as late as 2.8s after the onset of the adjective (1.5s after the offset of the noun). At first glance, this could be considered in conflict with previous work showing that semantic composition begins as early as 140ms after the onset of the noun, and disappears by 600ms (Bemis and Pylkkänen, 2013b). In our experiments, this would correspond to 0.94s and 1.4s after the onset of the adjective. Bemis and Pylkkänen contrasted two conditions: a pseudo word followed by a noun (no composition required) and adjective noun phrases. Thus, the early effects reported in previous work are, in some sense, the switching on of the machinery required to perform semantic composition, but not necessarily the time when we would expect to see the final product of semantic composition. If we compare the timing of the previous work to Figure 5.2a (decoding if a phrase begins with the word "the") we see two significantly above chance peaks centered near 1s and 1.25s (0.2-0.45s after the onset of the noun), which is within the range identified in previous studies. Perhaps the neural machinery that is not engaged for the pseudo word noun condition in Bemis and Pylkkänen (2013b) is also not engaged for the composition of determiners ("the") with nouns, for which composition is not necessary, or is considerably less.

There is some support for semantic effects as late as the effects we see here. Marinkovic et al. (2011) showed effects in MEG recordings during joke comprehension as late as 1.1s after the onset of the final word. Bastiaansen et al. (2010) found semantic violation effects in MEG signals as late as 2.5s after the onset of the critical word in the sentence. DeLong et al. (2014) varied the semantic plausibility of sentence critical words and found differences between the EEG recordings for anomalous and expected words that extended to the edge of their 1.2s analysis window (and possibly beyond). Many analyses have restricted themselves to the time period ending 1s after the onset of the critical word, possibly because the original windows for semantic and syntactic effects (at 400ms and 600ms respectively) extended only that far (Kutas and Hillyard, 1980; Kuperberg, 2007). The results of this study show that analyzing the signal beyond 1s could reveal new insight into semantic processing.

When we look for off-diagonal patches in our ROI analysis (Figure 5.7), we see the largest contributions come from occipital, limbic and parietal regions. Note that, although they have been implicated in previous studies of composition (Bemis and Pylkkänen, 2011, 2013b), temporal and IFG regions make only small contributions to the off-diagonal accuracy. Again, this previous work compared brain activation for composing vs not composing words, rather than for the final output of semantic composition.

One might question the results for the occipital ROI, as it is typically thought of as the locus of visual perception. However, previous work has found many semantic properties encoded in the occipital region of the brain (see supplementary material for (Sudre et al., 2012)). Many of the features decodable in the occipital region are visual (e.g. word length and verticality of picture stimuli), and in general, visual features are most decodable in occipital cortex before 200ms post stimulus onset. Note that in the Occipital plot in Figure 5.7, all off-diagonal activity appears at time periods 200ms and later, so it is likely not attributable to a visual feature of the stimuli being recalled. Additionally, Sudre et al. (2012) found that occipital cortex encoded more semantic features than any other ROI. Most of the semantic features encoded in occipital cortex related to the size, manipulability, animacy and threatening nature of the stimuli, with a few related to edibility[3]. These semantic features are highly related the attributes we manipulated with our adjective choices. Limbic areas code for features related to animacy, and parietal areas encode features related to size and animacy.

Recall, also, that we are using a corpus-derived semantic representation of the adjective for the decoding task in Figure 5.7. Though there are some correlates to the perceptual features of word strings in these corpus-derived features (e.g. determiners are often very short; frequent words are, on average, shorter than infrequent words) we are, by and large, decoding the semantic features of the words when we use these vectors.

5.7.3 Noun Semantics

The TTM for noun semantic decodability (Figure 5.6a) shows that the noun is decodable from just after noun onset until about 1.9s (1.1s after the onset of the noun). The noun's encoding is much less time-stable, as indicated by the thinness of the significantly above chance diagonal stripe in Figure 5.6a. During the time the adjective is presented, adjective decodability is consistent enough that TTM coordinates corresponding to time windows as far away as 400 ms have significantly above chance decodability. In comparison, noun decodability is much more time-localized, having a thickness (i.e. consistency in time) of about 200ms. This implies that the neural representation of the noun changes considerably over the time period that the noun is decodable. Sudre et al. (2012) showed that the semantics of a noun unfold as time progresses, perhaps also leading to a noun representation that is less stable in time. It is possible that there is more information to be recalled when processing a noun, which have many attribute-value pairs, rather than the few attribute-value pairs for the adjectives in this study.

Noun semantics are not decodable during the late decodability period of the adjective. It is somewhat counter-intuitive that the semantics of the adjective should be more salient than the

[3]For a full list of semantic features by ROI, see `http://www.cs.cmu.edu/~fmri/neuroimage2012_files/STScores.html`

semantics of the noun during the contemplation of the phrase. Perhaps this is an effect stemming from our choice of adjectives, which manipulate the most decodable features of the noun to their extreme ends.

Noun semantics also have very little in the way of significantly below chance decoding (Figure 5.6b), compared to the large significantly below chance patch present for adjective decoding during the presentation of the noun (Figure 5.5b). This implies that whatever change in the pattern that encodes the adjective's semantic code during the reading of the noun is not seen for the noun's semantic code - possibly because no word follows the noun. Future work could explore multi-word noun phrases (e.g. adjective-adjective-noun or noun-noun) to verify if each word is put into some "below chance" decoding state during the reading of the next word.

5.7.4 Significantly Below Chance Decoding in Train Time Matrices

Recall the pattern of significantly below chance 2 vs. 2 accuracy in the TTM for adjective semantics in Figure 5.5b. Significantly below chance decoding may seem counter-intuitive; how can the framework's predictions be systematically *worse* than random guessing? Let's work backwards from our decision function. Recall that the correct 2 vs. 2 choice satisfies

$$d(\mathbf{s_i}, \hat{\mathbf{s}}_\mathbf{i}) + d(\mathbf{s_j}, \hat{\mathbf{s}}_\mathbf{j}) < d(\mathbf{s_i}, \hat{\mathbf{s}}_\mathbf{j}) + d(\mathbf{s_j}, \hat{\mathbf{s}}_\mathbf{i})$$

Substituting in the definition of our distance function, we have:

$$\begin{gathered} 1 - cos(\mathbf{s_i}, \hat{\mathbf{s}}_\mathbf{i}) + 1 - cos(\mathbf{s_j}, \hat{\mathbf{s}}_\mathbf{j}) < 1 - cos(\mathbf{s_i}, \hat{\mathbf{s}}_\mathbf{j}) + 1 - cos(\mathbf{s_j}, \hat{\mathbf{s}}_\mathbf{i}) \\ -1(cos(\mathbf{s_i}, \hat{\mathbf{s}}_\mathbf{i}) + cos(\mathbf{s_j}, \hat{\mathbf{s}}_\mathbf{j})) < -1(cos(\mathbf{s_i}, \hat{\mathbf{s}}_\mathbf{j}) + cos(\mathbf{s_j}, \hat{\mathbf{s}}_\mathbf{i}) \\ cos(\mathbf{s_i}, \hat{\mathbf{s}}_\mathbf{i}) + cos(\mathbf{s_j}, \hat{\mathbf{s}}_\mathbf{j}) > cos(\mathbf{s_i}, \hat{\mathbf{s}}_\mathbf{j}) + cos(\mathbf{s_j}, \hat{\mathbf{s}}_\mathbf{i}) \end{gathered}$$

Where $cos(\mathbf{s_i}, \hat{\mathbf{s}}_\mathbf{i})$ is to the cosine of the angle between vectors s_i and $\hat{s}_i$. For predictions to be systematically wrong, we would require

$$cos(\mathbf{s_i}, \hat{\mathbf{s}}_\mathbf{i}) + cos(\mathbf{s_j}, \hat{\mathbf{s}}_\mathbf{j}) < cos(\mathbf{s_i}, \hat{\mathbf{s}}_\mathbf{j}) + cos(\mathbf{s_j}, \hat{\mathbf{s}}_\mathbf{i}).$$

One way to get this systematic flip of the comparison would be to multiply both sides of the inequality by -1, which would in turn require each of the cosine angles to be flipped. Cosine is negated when the angle differs by exactly π. In our scenario the true semantic vectors ($\mathbf{s_i}$ and $\mathbf{s_j}$) are fixed, so a negated cosine value would require that the predicted vectors point in the opposite direction (i.e. all coordinates are negated). Regressors trained on one time point could produce exactly negated predictions if the features they are tested on are themselves negated. Referring back to the definition of the regressor (Equation 5.1), if we set $x_j^{(i)} = -x_j^{(i)}$ for all j, then $\hat{s}^{(i)} = -\hat{s}^{(i)}$.

Note that if we negate the test data, the semantic vector will be negated, as will the cosine of the angles, and the 2 vs. 2 prediction. The change in 2 vs. 2 performance follows directly from the properties of the comparison function, and does not necessarily imply anything about the data. Instead, we test if the negation of the signal gives a better fit between time windows,

and how the fit varies based on the 2 vs. 2 accuracy of the corresponding TTM coordinate. We split the 2 vs. 2 accuracies into three bins: < 0.3, $0.3 - 0.7$ and > 0.7. For TTM coordinates $T_{i,j}$ in each of these ranges, we calculate the Mean Squared Error (MSE) between the corresponding MEG signals in windows i and j. MSE for the MEG signal $\mathbf{m} = \{m_1 \dots m_P\}$ at time points i and j is

$$\text{MSE}(i,j) = \frac{1}{P}\sum_{p=1}^{P}(m_p^{(i)} - m_p^{(j)})^2. \tag{5.6}$$

Lower MSE corresponds to a better fit between signals.

MSE results appear in Table 5.3. In Table 5.3 we see the MSE of the MEG signals is as expected: smaller when the 2 vs. 2 accuracy is higher. However, when we negate the signal from one window and recompute the MSE, we see that indeed, coordinates with low 2 vs. 2 accuracy have lower MSE (i.e. better fit), coordinates with 2 vs. 2 accuracy in the middle range show a small increase in MSE, and coordinates with high accuracy see a large rise in MSE. Thus the negated signal is a better fit in coordinates with low 2 vs. 2 accuracy, though not as good a fit as the non-negated signal in coordinates with high 2 vs. 2 accuracy. This asymmetry is due to the asymmetry in the distribution of the 2 vs. 2 accuracies in the TTM matrix: the 2 vs. 2 accuracy range is $[0.1 \dots 1]$. Thus, high accuracy is higher than low accuracy low, and it is logical that the fit of a negated signal from a low TTM coordinate does not achieve the MSE of a non-negated signal from a TTM coordinate with high accuracy.

We can also ask, in TTM coordinates with low 2 vs. 2 accuracy, are the predicted semantic features actually the negation of the true? To answer this question, for each training window (t_{train}), we select the window with highest (t_h) and lowest (t_ℓ) 2 vs. 2 accuracy, such that $h \neq train$ and $\ell \neq train$. Then, for each 2 vs. 2 test, we compare the 500 predicted semantic features from the window t_h and t_ℓ. We find that on average, 43% (about 215) of the features predicted from window t_ℓ have opposite sign compared to features from t_h. For a semantic features with opposite signs, the probability that the sign in t_ℓ is also negated with respect to the true semantic feature's sign is 64%. If the assignment of opposing signs (e.g. t_ℓ positive and t_h negative) were chosen via an independent random Bernoulli trial with equal probability for t_ℓ being either positive or negative, the chance of 64% of the 215 assigned features being flipped in a particular direction (i.e. so that the sign of t_ℓ disagrees with the true value) is $< 10^{-4}$. Thus this flipping of sign is likely not by chance, but a systematic effect caused by changes in the underlying MEG data.

What does it mean for the MEG signal to be negated? The MEG machine has two types of sensors, magnetometers and gradiometers. Magnetometers measure the direction of the magnetic field in a particular location, gradiometers measure the gradient of the magnetic field over space. In each of the 102 locations on the MEG helmet there is one magnetometer and two gradiometers. The two gradiometers in each location measure the gradient in perpendicular directions. MEG measures the post-synaptic potential of many parallel dendrites. Thus, the negation of the MEG signal equates to the reversal of the underlying magnetic field. This could be caused by several phenomena, one of which could be neurons leading to and from a location firing in alternation. Because our data is normalized, a negative feature value could also arise from a MEG sensor-time point that never actually produced a negative reading, but was shifted when the μ for that

Table 5.3: Mean Squared Error (MSE) of the MEG signals from two different time windows, partitioned based on 2 vs. 2 accuracy of the TTMs . MSE are calculated using the original MEG signals (column 2), or with the signal from one time window negated (column 3). Results are averaged over all 9 subjects.

2 vs. 2 Accuracy Range	MSE	MSE with one signal negated	Difference
< 0.3	2.00	1.90	0.10
$0.3 - 0.7$	1.91	1.98	-0.07
> 0.7	1.55	2.34	-0.79

sensor-time point was subtracted off.

For a more qualitative analysis of systematically low performance, we can examine the data from time windows where the off-diagonal 2 vs. 2 accuracy is low. Let's consider the occipital ROI for subject C during time windows 0.24-0.34s (during adjective presentation) and 1.40-1.50s (just after noun presentation). Figure 5.11 shows the data (X) for the phrase "light hammer" in the top row, weight matrix (β) in the leftmost column, and the product of data and weights ($X\beta$) in the interior cells of the table. Note that the sum of all elements in $X\beta$ equals the predicted value for that TTM coordinate for that phrase (shown in title of graph in each interior cell). β was trained to predict the first SVD dimension for adjective semantics. The true value in SVD dimension 1 for "light hammer" is -0.531, and predictions from misaligned time windows are both > 0.

Note that the incorrect predictions from the misaligned windows are driven largely by the mismatch of data and weights in the upper right hand corner, as well as the columns at time 0.04s. This is another piece of evidence that the code for adjectives is actually negated in the neural encoding, leading to the oscillatory patterns in the TTMs.

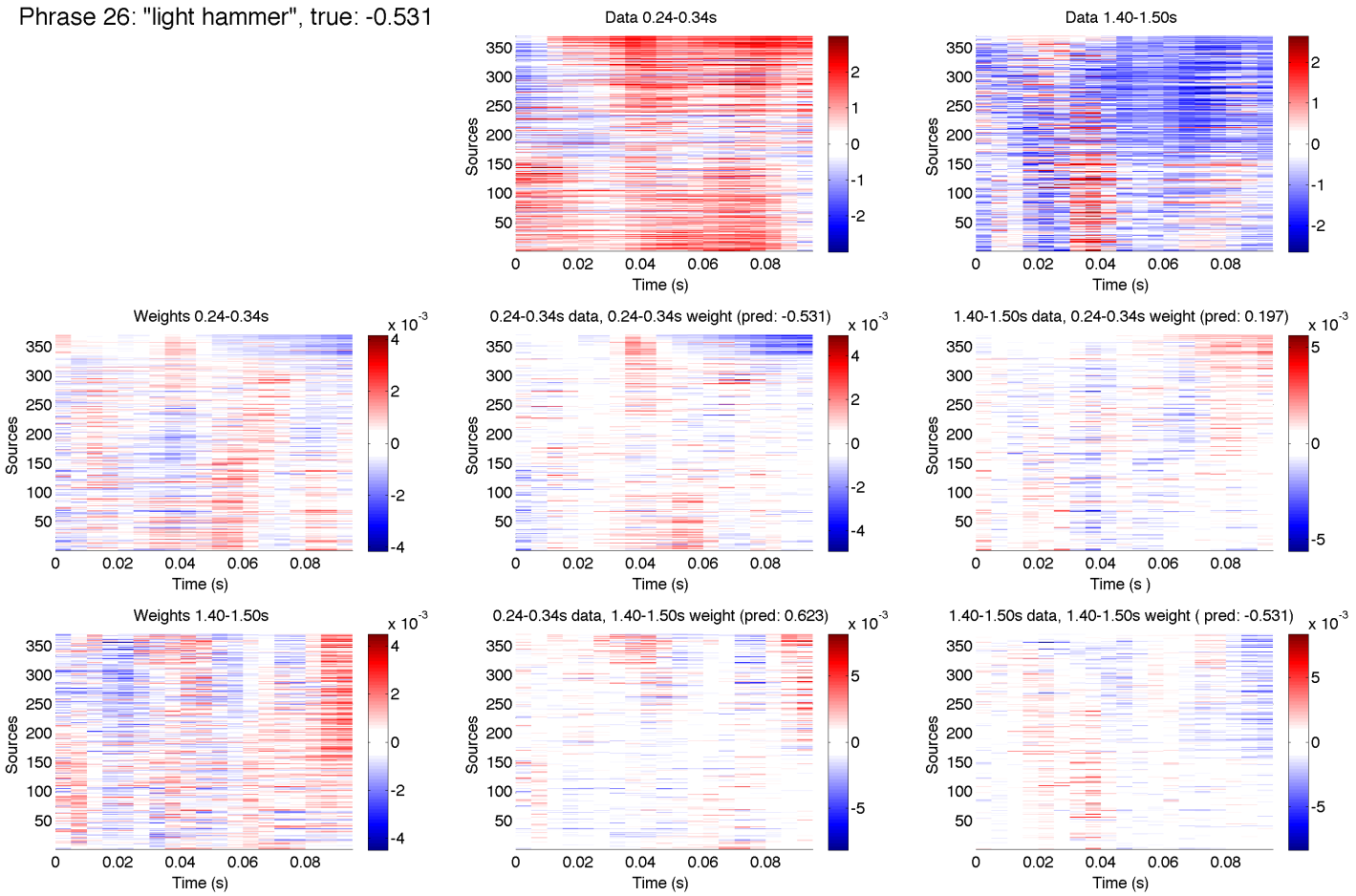
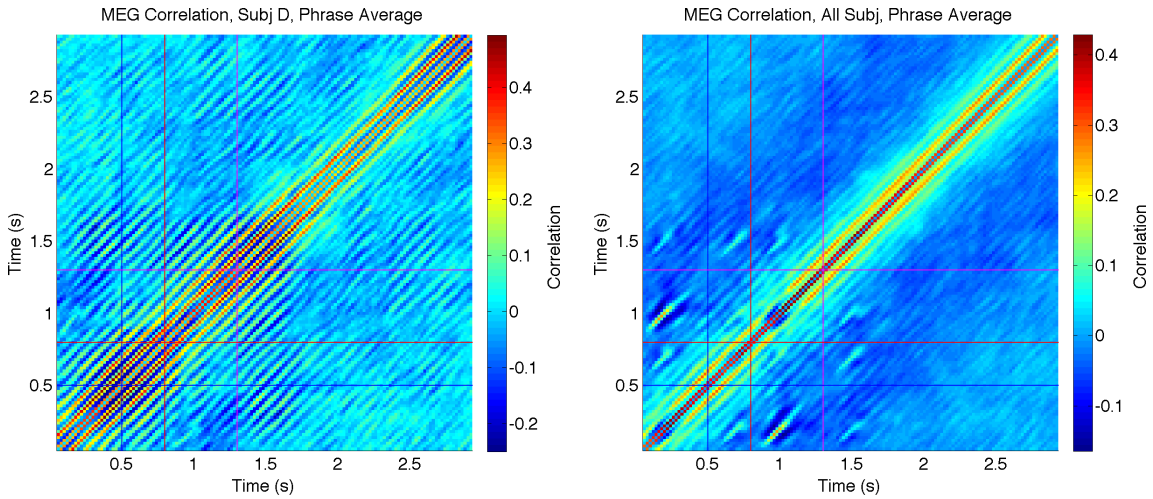


Figure 5.11: MEG data and trained weight matrices for two time windows (occipital ROI, subject C). Weight matrices were trained to predict the first SVD dimension of adjective semantic vectors. Weights/data points above zeros are red, below zero are blue. Top row: data from the ROI at two different time windows for the phrase "light hammer". Leftmost column: trained weight matrices for the two time windows. Interior of table: the product of the weight matrix (row) and the data (column), such that the sum of all elements of the matrix equals the predicted value (shown in the title of interior cells). The first SVD dimension for adjective "light" has value -0.531. Note that the incorrect predictions are driven largely by the mismatch of data and weights in the upper right hand corner, as well as the columns at time 0.04s.



(a) The correlation of the MEG signal for subject D, as a function of time.

(b) The correlation of the MEG signal for all subjects, as a function of time.

Figure 5.12: The correlation of the MEG signal as a function of time for all subjects and subject D. Correlation is calculated over all phrases within an 100ms window of MEG signal.

5.7.5 The Oscillatory Nature of Decodability

One of the most striking patterns in the TTMs is the oscillatory nature of decodability. In Figure 5.4 we can clearly see peaks and troughs in the decodability of the adjective. These oscillations align with the diagonal and repeat at approximately 10 Hz, within the typical alpha band of brain oscillations. Originally, alpha band oscillations were thought to be the brain's natural idling state, as they are observed in visual cortex when subjects close their eyes (Pfurtscheller et al., 1996). However, recent work has implicated the alpha band for a variety of active tasks. For example, alpha band activity has been implicated for short term memory tasks, increasing as the memory load increases (Jensen et al., 2002), and a variety of attentional tasks (Foxe and Snyder, 2011). It has been argued that alpha band activity is associated with access to the "knowledge system" and semantic orientation – neural processes proposed to be involved in understanding the world and one's relation to it in time, space and the entities in the vicinity (Klimesch, 2012).

Is it possible that the alpha-aligned decodability we observe is somehow related to a noise artifact? Each MEG session had 7 blocks (segments of continuous MEG recording with short breaks in between), so the chance of any noise artifact being aligned over all 7 blocks is very unlikely. In addition, the data was collected over a three month period, and while the decoding oscillations are less strong in some subjects they are clearly evident in the TTMs for all 9 subjects. There is also some subject (or session) variability to the oscillations. Some subjects' decoding oscillations are around 9Hz, and some are over 10 Hz. As a further point of evidence, Figure 5.12b shows the correlation of the average MEG signal over all 30 phrases and all subjects. If a noise artifact unrelated to brain activity was present, it should be revealed in this plot, but the strongest off-diagonal correlations are actually caused by the perceptual responses to the onset and offset of the stimuli (centered at around 0.2s, 0.7s, 1s and 1.5s). Compare the correlation matrix for subject D (Figure 5.12a) with the adjective semantics TTM for subject D

(Figure 5.4). Note that the pattern in both is still oscillatory, but the strongest correlation with the late MEG signal (2-2.5s) is with nearby time periods, not with time periods during the comprehension of the adjective (0-0.8s), as in the TTM. Note also the shift in the oscillations around 0.9-1s for subject D (Figure 5.4), which corresponds to the timing of the perceptual reaction to the onset of the noun stimulus. If this oscillatory pattern were indeed from an outside source, the fact that it occurs at exactly the expected onset of a perceptual reaction in the brain would be highly improbable. Together, these points are strong evidence that the oscillatory nature of word decoding is truly a brain-related signal, and not some noise superimposed on the signal itself.

5.7.6 Phrase Decoding

Phrasal encoding makes a brief appearance at 2s after the onset of the adjective (See Figures 5.9 and 5.10). It appears that subject variability may play a role in the timing of semantic composition (See Section 5.6.2), but further research will be required to verify that this is the case.

The noun-specific adjectives we chose for this experiment modulate some of the most salient features of the nouns (edibility, manipulability and animacy), as determined by previous studies of noun semantics (Sudre et al., 2012). In addition, the noun-specific adjectives are intersective, meaning that the semantics of the adjective are only minimally affected by the semantics of the noun. We believe these factors impacted our ability to decode phrasal meaning, but gave us the ability to detect adjective semantics late in time.

5.8 A Theory for Adjective Noun Composition

This chapter conveyed several new findings regarding adjective noun composition in the human brain:

- Adjective semantics are decodable for an extended period of time, continuously until 1.6s after the onset of the adjective. The neural encoding of adjective semantics is most stable from 0.2-0.65s after the onset of the adjective. The representation after 0.65 seconds is unstable and does not match to the neural encoding over 200ms away. This unstable adjective encoding period (0.65-1.6s) overlaps with noun presentation.
- Phrasal semantics are encoded 2s after the onset of the adjective (1.2s after the onset of the noun). This 2s peak is also seen for adjective semantic decoding and the late-onset timing is supported by previous research on higher-order semantics (Bastiaansen et al., 2010; Marinkovic et al., 2011; DeLong et al., 2014)
- Adjective semantics are reactivated during late processing, 2-3s after the onset of the adjective (1.2-2.2s after the onset of the noun). The reactivated encoding matches the stable representation seen during the initial 0.2-0.65s encoding.
- We hypothesize that the adjective representation is muted (but not abolished) during this period so that the noun semantics can be processed. In its muted form, it seems to best match the *negation* of the form seen during the stable 0.2-0.65s period.
- The noun's semantic representation is decodable from 1-1.9s. Noun encoding is unstable, and at no point does the encoding match an encoding more than 200ms away.

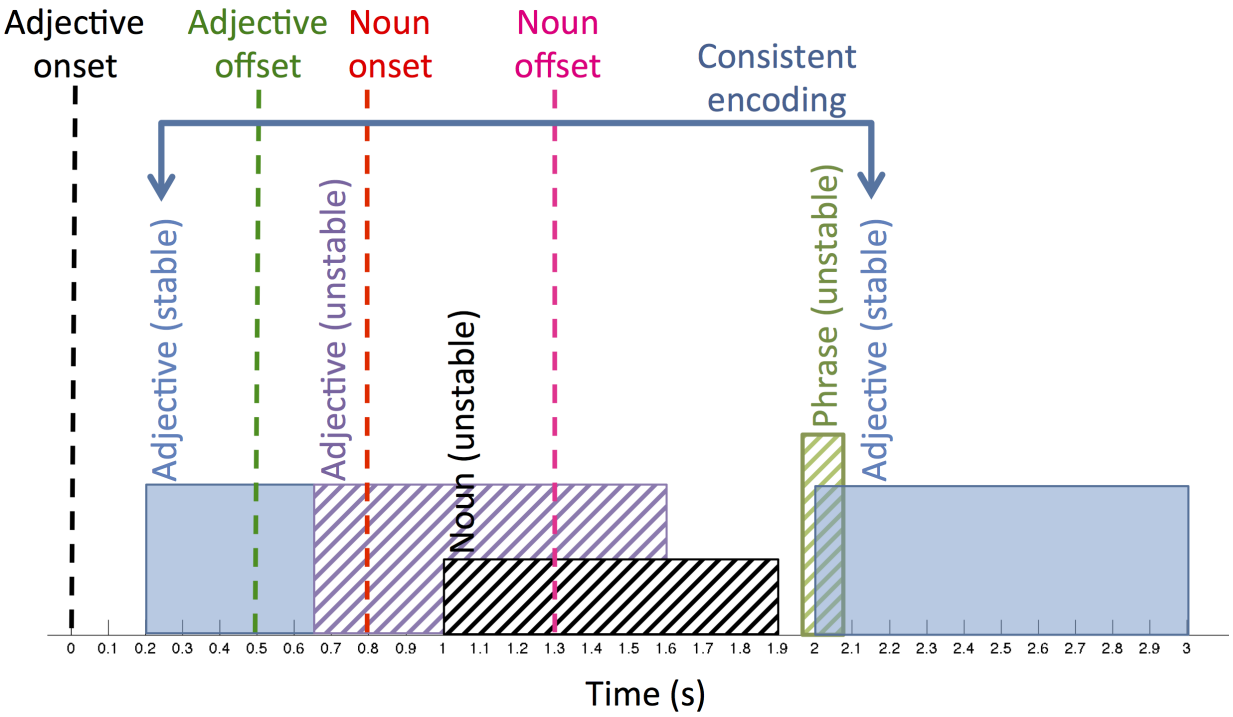


Figure 5.13: Decodability as a function of time for adjective, noun and phrasal semantics, based on the results of this study. The placement and width of rectangles indicates the onset and duration of a particular representation. Solid rectangles depict stable encodings, hatched depict unstable. The height of rectangles is just for visual contrast. Adjective encoding from 0.2-0.65s is stable, and is recalled in a consistent form during the period 2-3s. Adjective semantics are maintained until 1.6s, but in an unstable form after 0.65s. Noun semantics are encoded 1-1.9s, but are unstable during this period. Phrasal semantics make a brief appearance at 2s; the brevity may be due to the salience of the adjective encoding that resurges around 2s.

- Semantic encodings are oscillatory and alpha-aligned, as evidenced by the strong diagonal peaks and troughs of decodability in the TTMS.

Together, these findings paint a picture of the recall of adjectives and nouns in the brain, and the timing of some of the neural processing that takes place during adjective noun composition (summarized in Figure 5.13).

We theorize that adjective semantics are first recalled with a representation that is highly stable in time, but that is effectively stored away during the reading of the next word. In its stored form, the adjective can still be decoded, but the encoding is no longer time stable. While the adjective is in its stored form, noun semantics are retrieved. However, we hypothesize that nouns have a more multi-faceted semantic representation that is recalled in pieces over time (Sudre et al., 2012), leading to an unstable noun representation. About 1s after the noun stimuli onset the noun's semantics vanish and are replaced by phrasal semantics. This is followed by a resurgence of the adjective's initial stable representation. Because they so perfectly match the adjective encoding during adjective presentation, we believe this resurgence may not a property of the intersective adjectives chosen for this study, whose meaning is largely unaltered by composition with a noun. The onset of the output of adjective noun composition (phrasal semantics) is much later than the activation of the machinery responsible for combinatorics. Perhaps the combinatorial machinery acts like the conductor of an orchestra, and each semantic area is an instrument group. The conductor sends signals to the distributed areas to raise or lower their volume, or can ask specific areas to begin to play in synchrony. The combinatorial machinery is the hub that coordinates areas, readying them for the compositional processing.

5.9 Conclusion

In this chapter, we explored the semantic composition of adjective noun phrases in the human brain using the decodability of word properties as a function of time. We have found that the adjective and noun semantics are recalled during adjective and noun presentation respectively, but also that adjective semantics are held in mind during the time the noun is read and processed. At 1.2s post noun onset, we can decode the composed phrasal meaning, followed by a period of adjective semantic recall that is surprisingly consistent with the neural representation during adjective presentation. Neural semantic representations appear to be oscillatory and alpha-aligned.

We have found that the timing of phrasal semantics is much later than the findings of previous studies of adjective noun composition (Bemis and Pylkkänen, 2011, 2013a). However, we were in search of the *output* of semantic composition - the phrasal representation. In contrast, previous studies aimed to find the neural signs of the onset of compositional processes. Our results imply that future research interested in the composed representation consider should look beyond the typical 1s time window after the onset of a word.

With respect to semantic composition in the brain, there are many unanswered questions. For example, how would these results differ for different phrase types? How might one use these analysis techniques to explore sentential meaning, paragraph themes, and beyond? By exploring simple composition in a controlled setting, this study attempted to lay the groundwork for such future research directions. We hope our insights will aid in the exploration of semantics beyond simple adjective noun phrases.

Chapter 6

Discussion

This thesis brought together advances from the fields of machine learning, computational linguistics and psycholinguistics to further the study of semantics and semantic composition. We explored composition both abstractly, using patterns of word usage in a large text corpus, and in a grounded way, with recordings of brain activity. We showed that the types of semantic information available in brain images and corpus data are consistent and complementary.

This thesis shed new light on the neural representation of adjective noun phrases, and how that representation evolves over the sequential reading of words. The findings of this thesis are consistent with past work, but we also sought to answer fundamentally different questions: where and when in the brain can we find composed phrasal representations? And how does the neural encoding of words change as a function of time as reading progresses? We answered these questions, and in the process raised several new new research questions.

6.1 Summary of Contributions

This thesis made several contributions to the study of semantics using both corpora and brain imaging data.

Compositional constraints improve corpus-based models of composition. In Chapter 3 we developed an algorithm to learn a latent representation of semantics that incorporates the notion of semantic composition. Our model outperforms previous compositional models on several tasks, from more accurately predicting corpus statistics for adjective noun phrases, to better correlating with human judgements of phrase similarity. Human evaluators also judged the interpretable semantic representations from our model to be more consistent with phrasal meaning. We used the interpretability of our model to explore failure cases from two semantic composition tasks. This in-depth analysis allowed us to identify the shortcomings of our model. In particular, we found that the collision of multiple word senses in a single semantic representation can interfere with compositional methods.

Brain- and corpus-based models of semantics are consistent and complementary In Chapter 4 we presented JNNSE, which extended a previous matrix factorization algorithm, NNSE, to

incorporate an additional measure of semantics: brain activation data recorded while people read concrete nouns. Though we had brain image data for just a small number of words, we were able to show a positive effect on the learned latent representation of the model. When compared to a model that uses only one input data source, our joint model is more correlated to behavioral judgements of word semantics, can more accurately predict the corpus statistics for held out words, and can be used to predict words from a different person's brain imaging data, even when data was collected with a different brain imaging technology.

Collecting brain imaging data is time consuming and expensive, especially when compared to the minimal cost of harvesting large amounts of text from the internet. Thus, brain imaging data will likely not replace text data as a source of semantic information. However, our results show that there is semantic information available in brain imaging data that either does not exist, or has not yet been leveraged in text data. Thus, brain imaging data should be considered as a test bed for semantic models; increased performance on a brain image dataset implies that some of the additional information present in brain images is represented by the new model.

JNNSE joins two different sources of semantic data into one unified model: brain imaging data and corpus data. Brain imaging data from a single person can be thought of as a measurement of the instantiation of a "ground truth" semantic model. Corpus data, on the other hand, is comprised of the output of many brains (and thus many different semantic models). When we calculate statistics over large collections of text written by many different people, we effectively average the semantic models of many people into a single estimate. By combining these two different measurements of semantic information, we believe we have produced a model that is a more faithful representation of a community's mental vocabulary.

The neural encoding of adjectives is more time-stable than the neural encoding for nouns. In Chapter 5 we used brain imaging data to study the evolution of semantic composition for adjective noun phrases. We defined the notion of stability in neural encodings: a neural encoding is considered stable between time windows t_i and t_j if, when training data is taken from t_i and test data from t_j ($i \neq j$), we can still predict the word with high accuracy. We found that a semantic representation for the adjectives is held in mind for an extended period of time, until 1.6s after the onset of the adjective. The neural encoding of adjective semantics is stable from 0.2-0.65s after adjective onset. The adjective encoding enters into a less stable state at 0.65s, before the onset of the noun at 0.8s. The semantic representation of the noun is decodable 1-1.9s (0.2-1.1s after the onset of the noun). During this time, the encoding of the noun is unstable, and the encoding at one time point cannot be used to decode at a time point more than 200ms away.

We theorize that adjective semantics may be more stable because, for the adjectives used in this experiment, the semantic attributes manipulated by each adjective are relatively few in number. For example, the adjective "tasty" modulates the gustatory appeal of a noun, and possibly its smell and color. When the adjective is first called into mind, its representation may be stable because it is fairly simple. At 0.65s the adjective becomes unstable, which we theorize is in preparation for the processing of the noun. In its unstable, "stored" form, the negation of the MEG signal seems to better match the form during the stable period, but this conjecture requires further investigation to confirm.

We theorize that the encoding of the noun may be unstable because nouns are composed of

many attribute value pairs, and thus are more complex than the adjectives we selected for this study. We believe that the instability of the noun's encoding could be due to this complexity, which could require the unfolding of semantics (attribute-value pairs) over time, as evidenced in previous studies (Sudre et al., 2012).

The neural encoding of composed phrasal semantics is available 1s post noun onset. We found the semantics of the phrase to be encoded much later than previous studies of semantic composition may have indicated. We see phrasal semantics encoded at 2s, 1.2s after the onset of the noun. This is much later than the onset combinatorial processing revealed by previous work, which has been shown to start as early as 0.2s after the onset of the noun. Our results are not incompatible with previous results, as we were searching for the final output of semantic composition (phrasal semantics), whereas previous research was searching for the *first* differences in the neural processing of a stimuli requiring composition.

We found there to be some subject-dependent variance for the onset of the encoding of phrasal semantics, though further research will be required to confirm this. Still, it is highly likely that, when it comes to studying more complex linguistic tasks like semantic composition, the variability of timing across subjects may become a more important factor.

The neural encoding for an adjective resurges after the neural encoding of the composed phrase. After the unstable representation of the adjective fades at 1.65s, there is a brief period (1.65-2s) where the adjective cannot be decoded at all. At 2s post adjective onset, we see a resurgence of the stable representation of the adjective originally seen at 0.2-0.65s. This late-onset neural representation of the adjective so closely matches the early encoding that we can train and test classifiers during these two distant time points and produce 2 vs. 2 accuracy significantly above chance. It is possible that this resurgence in adjective encoding may be a property of intersective adjectives, whose meaning is largely unaltered by composition with a noun.

The neural encoding of semantics is oscillatory and can lead to significantly below chance decodability The neural encoding of adjectives appears to be oscillatory, evidenced by the fact that the neural encodings best match when the distance between windows is a multiple of some constant. In our data, this constant equals the length of one alpha oscillation. When training and testing windows are exactly anti-phase (i.e. half of the time constant away), decodability can drop significantly below chance. We explored some possible explanations for this significantly below chance representation and found evidence for it being related to a negation of the pattern responsible for decodability. These findings have raised many questions, and we leave most of them for future work.

Composed phrase meaning exists outside of areas previously implicated in semantic composition. Much of the related work on semantic composition focused on identifying areas of the brain that are differentially engaged in compositional vs non-compositional tasks (Bemis and Pylkkänen, 2011, 2013a), or how the brain reacts differently to stimuli showing semantically or syntactically anomalous sentences (Kutas and Hillyard, 1980; Hagoort, 2005; Kuperberg, 2007) (see Section 2.2). Thus, the previous work answers a question that is separate (though related)

from the questions posed in this thesis. In this thesis we searched not for the areas involved in composing meaning, but for the timing and location of the *final* composed semantic representation, and how it differs as a function of the input stimuli. Our results are not incompatible with these previous studies, as we asked a fundamentally different research question.

One previous study which could be considered at odds with our work is Baron and Osherson (2011). Baron and Osherson studied the semantic composition of adjective noun phrases using fMRI and visual stimuli. The stimuli was faces of young or old males (boys and men) and young or old females (girls and women). In the fMRI scanner, the faces were presented in blocks. For each block within the experiment, subjects were given a category (e.g. girl) and asked determine if each of the stimuli faces was a member of that category. Thus, for each block the face stimuli were the same, and only the concept being matched differed. Thus, any differences in activation can be attributed only to the matching task, and not to the stimuli. Baron and Osherson then created conceptual maps by learning regressors to predict brain activity based on the age (young or old) and gender of the matching task. They found that the conceptual maps of the adjective and noun (e.g. the map of young and the map of male) could be added or multiplied to approximate the activation for the composed category (e.g. boy). They found that the areas of the brain that could be approximated well with an additive function were widespread, whereas the multiplicative function localized just to the left anterior temporal lobe (LATL).

At first glance, the results of Baron and Osherson (2011) may seem to imply that the composed semantic meaning for the concepts resides only in the LATL. However, the results actually imply that LATL is the only area *shared* between age and gender related concepts during their composition, and thus is the only area to survive the multiplication of brain activation patterns. So, while the compositional activity in LATL can be predicted by the multiplication (or addition) of the activity for each of the constituent concepts, the widespread activation outside of LATL can be approximated best with addition. This implies that the distributed areas of activation for concepts are disjoint. We argue that if the composed concepts reside in disjoint brain areas, disjoint areas will be involved in semantic composition. Requiring the overlap of brain areas can identify areas of the brain that code for both concepts *or* that coordinate disjoint areas during composition. In the case of the coordination of disjoint areas, the multiplicative map is predictive only if coordination produces a pattern that depends on the areas being coordinated. We theorize this coordinating pattern in LATL could stem from differing connectivity to different distributed areas during composition. Thus, our results are consistent with both Baron and Osherson (2011) and Bemis and Pylkkänen (2011, 2013a) and may actually provide a piece of unifying evidence that the result of semantic composition is widespread, but the machinery shared over compositional tasks resides, at least partially, in LATL.

This thesis also provides evidence that final composed semantic representation may not be encoded neurally until 1s after the onset of the final stimuli. This finding may encourage future research to record activity beyond 1s, and consider analyzing later time periods.

6.2 Future work

There are several new directions possible from the results of this thesis. This thesis showed that corpus and brain data are complimentary sources of semantic information. There are many other

sources of semantic information (e.g. behavioral data, image data), and each of these could be incorporated into semantic models to create a more complete semantic picture. Indeed some studies have already headed in this direction (Bruni et al., 2011; Silberer and Lapata, 2012; Silberer et al., 2013) Children learn language in a very multi-modal way, using all of their senses and their ability to interact with the environment. If we use language acquisition as inspiration, using additional information sources like images is a natural extension to a semantic model.

Our adjective noun decoding results raise several questions ripe for future work. In particular the exploration of the oscillatory nature of semantic encodings could be very instructive. What do these oscillations say about the nature of semantic encoding? How do oscillations relate to the binding problem (Treisman, 1996), associating particular features with particular objects? For example, how can we simultaneously think of a black dog and a white cat, without blending the two to make a two grey house pets? The brain segregates semantic features into object-specific groups, and the phase-aligned timing of the neural encoding could be one way in which that is accomplished.

Significantly below chance decoding is another interesting finding from this thesis, and is probably closely associated with the oscillatory nature of semantic encoding. Future work could further analyze the pattern for encoding words in an anti-phase state. During the anti-phase state, the encodings become unstable. What factors in the encoding cause this onset of instability? How does the anti-phase state interact with the neural encoding of newly read words?

There are also many opportunities to extend this work to other types of phrases (e.g. noun noun phrases) or more complex linguistic stimuli like sentences and stories. Are the oscillatory patterns seen here also present in those scenarios? Does the anti-phase signature of neural encoding also appear for more complex stimuli? Is the resurgence of single word semantic encodings seen at the end of sentences?

This thesis has shown how brain imaging data and corpus data can be used together to build better models of semantics, and to better understand semantic representations in the brain. We have shown several ways in which machine learning algorithms can aid in the exploration of complex data sources like brain imaging and corpus data. We also offered new insights into the way the human brain performs the semantic composition of adjective noun phrases. We hope this work informs further research into the nature of semantic composition both in the brain, and more abstractly in patterns of word usage.

Appendix

A Adjective Noun Brain Imaging Materials

The phrases for the adjective noun brain imaging experiment are made from 6 nouns (“dog”, “bear”, “tomato”, “carrot”, “hammer”, “shovel”) and 8 adjectives (“big”, “small”, “ferocious”, “gentle”, “light”, “heavy”, “rotten”, ’‘tasty”), as well as two null words: “the” and “thing”. The phrases are:

- the dog
- the bear
- the tomato
- the carrot
- the hammer
- the shovel
- big dog
- big bear
- big tomato
- big carrot
- big hammer
- big shovel
- small dog
- small bear
- small tomato
- small carrot
- small hammer
- small shovel
- ferocious dog
- ferocious bear
- gentle dog
- gentle bear
- light hammer
- light shovel
- heavy hammer
- heavy shovel
- rotten carrot
- rotten tomato
- tasty carrot
- tasty tomato
- big thing
- small thing
- ferocious thing
- gentle thing
- light thing
- heavy thing
- rotten thing
- tasty thing

The phrases were presented in rapid serial visual presentation. Each word of the phrase appears by itself on the screen, with a fixation cross between the words of the phrase, as well as between phrases. Each word appears on the screen for 500ms, with a 300ms break between words of a phrase. There is 3s total time between the onset of the adjective in consecutive phrase presentations. Each phrase was presented 20 times, randomly ordered in 7 experimental blocks. The random order was chosen such that the same phrase was never repeated twice in a row. Between experiment blocks, subjects were given the option to take a break.

Oddball stimuli (two adjectives instead of an adjective noun phrase) were created by pairing adjectives. The word light was omitted from the oddball pairs because it has both an adjective an noun sense. Oddballs appeared after 10% of phrases, randomly inserted such that there was an equal number of oddballs per block. Two oddball phrases were never presented in succession.

The adjective noun phrase following an oddball trial was omitted from the analysis, to avoid any contamination of trials by movement artifacts.

Bibliography

Andrew J Anderson, Elia Bruni, Ulisse Bordignon, Massimo Poesio, and Marco Baroni. Of words , eyes and brains : Correlating image-based distributional semantic models with neural representations of concepts. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing*, 2013.

Sean G Baron. *The neural basis of compositionality: Functional magnetic resonance imaging studies of conceptual combination*. PhD thesis, Princeton, 2012.

Sean G Baron and Daniel Osherson. Evidence for conceptual combination in the left anterior temporal lobe. *NeuroImage*, 55(4):1847–52, April 2011. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2011.01.066.

Marco Baroni and Alessandro Lenci. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721, 2010.

Marco Baroni and Roberto 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*, pages 1183–1193. Association for Computational Linguistics, 2010.

Marco Baroni, Georgiana Dinu, and German Kruszewski. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *52nd Annual Meeting of the Association for Computational Linguistics*, 2014.

Marcel Bastiaansen, Lilla Magyari, and Peter Hagoort. Syntactic unification operations are reflected in oscillatory dynamics during on-line sentence comprehension. *Journal of cognitive neuroscience*, 22(7):1333–47, July 2010. ISSN 1530-8898. doi: 10.1162/jocn.2009.21283.

Mark F Bear, Barry W Connors, and Michael A Paradiso. *Neuroscience: Exploring the brain (3rd ed.)*. 2007. ISBN 0-7817-6003-8 (Hardcover).

D K Bemis and L Pylkkänen. Basic linguistic composition recruits the left anterior temporal lobe and left angular gyrus during both listening and reading. *Cerebral cortex (New York, N.Y. : 1991)*, 23(8):1859–73, August 2013a. ISSN 1460-2199. doi: 10.1093/cercor/bhs170.

Douglas K Bemis and Liina Pylkkänen. Simple composition: a magnetoencephalography investigation into the comprehension of minimal linguistic phrases. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 31(8):2801–14, February 2011. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.5003-10.2011.

Douglas K Bemis and Liina Pylkkänen. Flexible composition: MEG evidence for the deployment

of basic combinatorial linguistic mechanisms in response to task demands. *PloS one*, 8(9): e73949, January 2013b. ISSN 1932-6203. doi: 10.1371/journal.pone.0073949.

William Blacoe and Mirella Lapata. A Comparison of Vector-based Representations for Semantic Composition. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 546–556, Jeju Island, Korea, 2012.

David M Blei and Jon D. McAuliffe. Supervised topic models. In *Advances in Neural Information Processing Systems*, pages 1–22, 2007.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022, 2003. ISSN 15324435. doi: 10.1162/jmlr.2003.3.4-5.993. URL `http://www.crossref.org/jmlr_DOI.html`.

Stephen Boyd. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2010. ISSN 1935-8237. doi: 10.1561/2200000016.

Elia Bruni, Giang Binh Tran, and Marco Baroni. Distributional semantics from text and images. In *Proceedings of the EMNLP 2011 Geometrical Models for Natural Language Semantics (GEMS)*, 2011.

John A Bullinaria and Joseph P Levy. Limiting factors for mapping corpus-based semantic representations to brain activity. *PloS one*, 8(3):e57191, January 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0057191.

Jamie Callan and Mark Hoy. The ClueWeb09 Dataset, 2009. URL `http://boston.lti.cs.cmu.edu/Data/clueweb09/`.

Alexander M Chan, Eric Halgren, Ksenija Marinkovic, and Sydney S Cash. Decoding word and category-specific spatiotemporal representations from MEG and EEG. *NeuroImage*, 54(4): 3028–39, February 2011. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2010.10.073.

Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M Blei. Reading Tea Leaves : How Humans Interpret Topic Models. In *Advances in Neural Information Processing Systems*, pages 1–9, 2009a.

Kai-min 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 Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, pages 638–646, 2009b.

Kai-min Kevin Chang, Tom Mitchell, and Marcel Adam Just. Quantitative modeling of the neural representation of objects: how semantic feature norms can account for fMRI activation. *NeuroImage*, 56(2):716–27, May 2011. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2010.04.271.

Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.

Bryan R Conroy, Benjamin D Singer, J Swaroop Guntupalli, Peter J Ramadge, and James V Haxby. Inter-subject alignment of human cortical anatomy using functional connectivity. *Neu-*

roImage, 81:400–11, November 2013. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2013.05.009.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

S Dejong. SIMPLS - AN ALTERNATIVE APPROACH TO PARTIAL LEAST-SQUARES REGRESSION. *Chemometrics and Intelligent Laboratory Systems*, 18:251–263, 1993. ISSN 01697439. doi: 10.1016/0169-7439(93)85002-x.

Katherine a DeLong, Laura Quante, and Marta Kutas. Predictability, plausibility, and two late ERP positivities during written sentence comprehension. *Neuropsychologia*, 61:150–62, August 2014. ISSN 1873-3514. doi: 10.1016/j.neuropsychologia.2014.06.016.

Georgiana Dinu, Nghia The Pham, and Marco Baroni. General estimation and evaluation of compositional distributional semantic models. In *Workshop on Continuous Vector Space Models and their Compositionality*, Sofia, Bulgaria, 2013.

Andrew D Engell, Scott Huettel, and Gregory McCarthy. The fMRI BOLD signal tracks electrophysiological spectral perturbations, not event-related potentials. *NeuroImage*, 59(3):2600–6, February 2012. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2011.08.079.

Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

John J Foxe and Adam C Snyder. The Role of Alpha-Band Brain Oscillations as a Sensory Suppression Mechanism during Selective Attention. *Frontiers in psychology*, 2(July):154, January 2011. ISSN 1664-1078. doi: 10.3389/fpsyg.2011.00154.

Alona Fyshe, Partha Talukdar, Brian Murphy, and Tom Mitchell. Documents and Dependencies : an Exploration of Vector Space Models for Semantic Composition. In *Computational Natural Language Learning*, Sofia, Bulgaria, 2013.

Alona Fyshe, Partha P Talukdar, Brian Murphy, and Tom M Mitchell. Interpretable Semantic Vectors from a Joint Model of Brain-and Text-Based Meaning. In *Association for Computational Linguistics*, Baltimore, MD, USA, 2014.

Alona Fyshe, Leila Wehbe, Partha P. Talukdar, Brian Murphy, and Tom M. Mitchell. A Compositional and Interpretable Semantic Space. In *Proceedings of the NAACL-HLT*, Denver, USA, May 2015. Association for Computational Linguistics.

Arthur M Glenberg and David a Robertson. Symbol Grounding and Meaning: A Comparison of High-Dimensional and Embodied Theories of Meaning. *Journal of Memory and Language*, 43(3):379–401, October 2000. ISSN 0749596X. doi: 10.1006/jmla.2000.2714.

E Grefenstette, G Dinu, YZ Zhang, M. Sadrzadeh, and M. Baroni. Multi-step regression learning for compositional distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*, 2013.

Sunil Kumar Gupta, Dinh Phung, Brett Adams, and Svetha Venkatesh. Regularized nonnegative shared subspace learning. *Data Mining and Knowledge Discovery*, 26(1):57–97, 2013. ISSN 1384-5810. doi: 10.1007/s10618-011-0244-8.

Peter Hagoort. On Broca, brain, and binding: a new framework. *Trends in cognitive sciences*, 9 (9):416–23, September 2005. ISSN 1364-6613. doi: 10.1016/j.tics.2005.07.004.

Peter Hagoort. Nodes and networks in the neural architecture for language: Broca's region and beyond. *Current opinion in neurobiology*, 28C:136–141, July 2014. ISSN 1873-6882. doi: 10.1016/j.conb.2014.07.013.

Emma L Hall, Siân E Robson, Peter G Morris, and Matthew J Brookes. The relationship between MEG and fMRI. *NeuroImage*, November 2013. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2013.11.005.

M. Hamalainen, R. Hari, R.J. Ilmoniemi, J. Knuutila, and O.V. Lounasmaa. Magnetoencephalographytheory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of modern Physics*, 65(2), 1993.

M S Hämäläinen and R J Ilmoniemi. Interpreting magnetic fields of the brain: minimum norm estimates. *Medical & biological engineering & computing*, 32:35–42, 1994. ISSN 0140-0118. doi: 10.1007/BF02512476.

Peter Hansen, Morten Kringelbach, and Riitta Salmelin. *MEG: An Introduction to Methods*. Oxford University Press, USA, 2010. ISBN 0195307232.

Karl Moritz Hermann, Edward Grefenstette, and Phil Blunsom. " Not not bad" is not" bad": A distributional account of negation. In *Proceedings of the ACL Workshop on Continuous Vector Space Models and their Compositionality*, Sofia, Bulgaria, 2013.

Gregory Hickok and David Poeppel. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition*, 92(1-2):67–99, 2004. ISSN 0010-0277. doi: 10.1016/j.cognition.2003.10.011.

Gregory Hickok and David Poeppel. The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(May):393–402, 2007.

John C J Hoeks, Laurie a Stowe, and Gina Doedens. Seeing words in context: the interaction of lexical and sentence level information during reading. *Cognitive brain research*, 19(1):59–73, March 2004. ISSN 0926-6410. doi: 10.1016/j.cogbrainres.2003.10.022.

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

Alexander G Huth, Shinji Nishimoto, An T Vu, and Jack L Gallant. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–24, December 2012. ISSN 1097-4199. doi: 10.1016/j.neuron.2012.10.014.

Chang Hwan Im, Arvind Gururajan, Nanyin Zhang, Wei Chen, and Bin He. Spatial resolution of EEG cortical source imaging revealed by localization of retinotopic organization in human primary visual cortex. *Journal of Neuroscience Methods*, 161:142–154, 2007. ISSN 01650270. doi: 10.1016/j.jneumeth.2006.10.008.

Ole Jensen, Jack Gelfand, John Kounios, and JE Lisman. Oscillations in the alpha band (912 Hz) increase with memory load during retention in a short-term memory task. *Cerebral Cortex*, 12 (8):877–882, 2002.

Yangqing Jia and Trevor Darrell. Factorized Latent Spaces with Structured Sparsity. In *Advances*

in Neural Information Processing Systems, volume 23, 2010.

Marcel Adam Just, Vladimir L Cherkassky, Sandesh Aryal, and Tom M Mitchell. A neurosemantic theory of concrete noun representation based on the underlying brain codes. *PloS one*, 5(1):e8622, January 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0008622.

Wolfgang Klimesch. -Band Oscillations, Attention, and Controlled Access To Stored Information. *Trends in cognitive sciences*, 16(12):606–17, December 2012. ISSN 1879-307X. doi: 10.1016/j.tics.2012.10.007.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2 (November):4, January 2008a. ISSN 1662-5137. doi: 10.3389/neuro.06.004.2008.

Nikolaus Kriegeskorte, Marieke Mur, Douglas A Ruff, Roozbeh Kiani, Jerzy Bodurka, Hossein Esteky, Keiji Tanaka, and Peter A Bandettin. Matching Categorical Object Representations in Inferior Temporal Cortex of Man and Monkey. *Neuron*, 60(6):1126–1141, 2008b. doi: 10.1016/j.neuron.2008.10.043.Matching.

Jayant Krishnamurthy and Tom M Mitchell. Vector Space Semantic Parsing : A Framework for Compositional Vector Space Models. In *Workshop on Continuous Vector Space Models and their Compositionality*, Sofia, Bulgaria, 2013.

Gina R Kuperberg. Neural mechanisms of language comprehension: challenges to syntax. *Brain research*, 1146:23–49, May 2007. ISSN 0006-8993. doi: 10.1016/j.brainres.2006.12.063.

M Kutas and SA Hillyard. Reading senseless sentences: brain potentials reflect semantic incongruity. *Science*, 207(4427):203–5, 1980.

T Landauer and S Dumais. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2): 211–240, 1997a.

TK Landauer and ST Dumais. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 1(2):211–240, 1997b.

TK Landauer and Darrell Laham. How Well Can Passage Meaning be Derived without Using Word Order? A Comparison of Latent Semantic Analysis and Humans. In *Proceedings of the 19th annual meeting of the Cognitive Science Society*, pages 412–417, 1997.

Omer Levy and Yoav Goldberg. Neural Word Embedding as Implicit Matrix Factorization. In *Advances in Neural Information Processing Systems*, pages 1–9, 2014.

Dekang Lin. Automatic Retrieval and Clustering of Similar Words. In *Proceedings of COLING-ACL*, pages 768–774, 1998.

K Lund and C Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, and Computers*, 28:203–208, 1996a.

Kevin Lund and Curt Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence, 1996b. ISSN 0743-3808.

Julien Mairal, Francis Bach, J Ponce, and Guillermo Sapiro. Online learning for matrix factor-

ization and sparse coding. *The Journal of Machine Learning Research*, 11:19–60, 2010.

Ksenija Marinkovic, Sharelle Baldwin, Maureen G Courtney, Thomas Witzel, Anders M Dale, and Eric Halgren. Right hemisphere has the last laugh: neural dynamics of joke appreciation. *Cognitive, affective & behavioral neuroscience*, 11(1):113–30, March 2011. ISSN 1531-135X. doi: 10.3758/s13415-010-0017-7.

Ken McRae, George S Cree, Mark S Seidenberg, and Chris McNorgan. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37 (4):547–59, November 2005. ISSN 1554-351X.

Tomá Mikolov. *Statistical Language Models based on Neural Networks*. Phd thesis, Brno University of Technology, 2012.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of Neural Information Processing Systems*, pages 1–9, 2013.

Jeff Mitchell and Mirella Lapata. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–429, November 2010. ISSN 1551-6709. doi: 10.1111/j.1551-6709.2010.01106.x.

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 (New York, N.Y.)*, 320(5880):1191–5, May 2008. ISSN 1095-9203. doi: 10.1126/science.1152876.

Brian Murphy, Partha Talukdar, and Tom Mitchell. Selecting Corpus-Semantic Models for Neurolinguistic Decoding. In *First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 114–123, Montreal, Quebec, Canada, 2012a.

Brian Murphy, Partha Talukdar, and Tom Mitchell. Learning Effective and Interpretable Semantic Models using Non-Negative Sparse Embedding. In *Proceedings of Conference on Computational Linguistics (COLING)*, 2012b.

Brian Murphy, Partha Pratim Talukdar, and Tom Mitchell. Learning Effective and Interpretable Semantic Models using Non-Negative Sparse Embedding. In *International Conference on Computational Linguistics (COLING 2012), Mumbai, India*, 2012c.

Andrew Y. Ng and Michael I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems*, volume 14, 2002.

S Padó and M Lapata. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199, 2007.

Mark Palatucci, Geoffrey Hinton, Dean Pomerleau, and Tom M Mitchell. Zero-Shot Learning with Semantic Output Codes. *Advances in Neural Information Processing Systems*, 22:1410–1418, 2009.

G. Pfurtscheller, a. Stancák, and Ch. Neuper. Event-related synchronization (ERS) in the alpha band an electrophysiological correlate of cortical idling: A review. *International Journal of Psychophysiology*, 24(1-2):39–46, November 1996. ISSN 01678760. doi:

10.1016/S0167-8760(96)00066-9.

Rajeev D S Raizada and Andrew C Connolly. What Makes Different People's Representations Alike : Neural Similarity Space Solves the Problem of Across-subject fMRI Decoding. *Journal of Cognitive Neuroscience*, 24(4):868–877, 2012.

Indrayana Rustandi, Marcel Adam Just, and Tom M Mitchell. Integrating Multiple-Study Multiple-Subject fMRI Datasets Using Canonical Correlation Analysis. In *MICCAI 2009 Workshop: Statistical modeling and detection issues in intra- and inter-subject functional MRI data analysis*, 2009.

Mehrnoosh Sadrzadeh and Edward Grefenstette. A Compositional Distributional Semantics Two Concrete Constructions and some Experimental Evaluations. *Lecture Notes in Computer Science*, 7052:35–47, 2011.

Magnus Sahlgren. *The Word-Space Model Using distributional analysis to represent syntagmatic and paradigmatic relations between words*. Doctor of philosophy, Stockholm University, 2006a.

Magnus Sahlgren. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Dissertation, Stockholm University, 2006b. URL `http://en.scientificcommons.org/20017188`.

Carina Silberer and Mirella Lapata. Grounded models of semantic representation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1423–1433, 2012.

Carina Silberer, Vittorio Ferrari, and Mirella Lapata. Models of Semantic Representation with Visual Attributes. In *Association for Computational Linguistics 2013*, Sofia, Bulgaria, 2013.

Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. Semantic Compositionality through Recursive Matrix-Vector Spaces. In *Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2012.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Conference on Empirical Methods on Natural Language Processing*, pages 1631–1642, 2013.

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(1):463–451, May 2012. ISSN 10538119. doi: 10.1016/j.neuroimage.2012.04.048.

S Taulu and J Simola. Spatiotemporal signal space separation method for rejecting nearby interference in MEG measurements. *Physics in Medicine and Biology*, 51:1–10, 2006. doi: 10.1088/0031-9155/51/0/000.

Samu Taulu and Riitta Hari. Removal of magnetoencephalographic artifacts with temporal signal-space separation: demonstration with single-trial auditory-evoked responses. *Human brain mapping*, 30(5):1524–34, May 2009. ISSN 1097-0193. doi: 10.1002/hbm.20627.

Samu Taulu, Matti Kajola, and Juha Simola. The Signal Space Separation method. *ArXiv Physics*, 2004.

Anne Treisman. The binding problem. *Current opinion in neurobiology*, 6(2):171–178, April 1996. ISSN 09594388. doi: 10.1016/S0959-4388(96)80070-5.

Peter D Turney. Domain and Function : A Dual-Space Model of Semantic Relations and Compositions. *Journal of Artificial Intelligence Research*, 44:533–585, 2012.

Peter D Turney and Patrick Pantel. From Frequency to Meaning : Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37:141–188, 2010.

M A Uusitalo and R J Ilmoniemi. Signal-space projection method for separating MEG or EEG into components. *Medical & biological engineering & computing*, 35(2):135–40, March 1997. ISSN 0140-0118.

Ashish Vaswani, Y Zhao, Victoria Fossum, and David Chiang. Decoding with Large-Scale Neural Language Models Improves Translation. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing (EMNLP)*, pages 1387–1392, 2013.

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

Martha White, Yaoliang Yu, Xinhua Zhang, and Dale Schuurmans. Convex multi-view subspace learning. In *Advances in Neural Information Processing Systems*, pages 1–14, 2012.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, April 2005. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2005.00503.x.