ABSTRACT

Title of Document: USING A HIGH-DIMENSIONAL MODEL OF SEMANTIC SPACE TO PREDICT NEURAL ACTIVITY.

Alice Freeman Jackson, Doctor of Philosophy, 2014

Directed By: Professor Donald J. Bolger, Department of Human Development and Quantitative Methodology

This dissertation research developed the GOLD model (Graph Of Language Distribution), a graph-structured semantic space model constructed based on co-occurrence in a large corpus of natural language, with the intent that it may be used to explore what information may be present about relationships between words in such a model and the degree to which this information may be used to predict brain responses and behavior in language tasks. The present study employed GOLD to examine genera relatedness as well as two specific types of relationship between words: semantic similarity, which refers to the degree of overlap in meaning between words, and associative relatedness, which refers to the degree to which two words occur in the same schematic context. It was hypothesized that this graph-structured model of language constructed based on co-occurrence should easily capture associative relatedness, because this type of relationship is thought to be present

directly in lexical co-occurrence. Additionally, it was hypothesized that semantic similarity may be extracted from the intersection of the set of first-order connections, because two words that are semantically similar may occupy similar thematic or syntactic roles across contexts and thus would co-occur lexically with the same set of nodes. Based on these hypotheses, a set of relationship metrics were extracted from the GOLD model, and machine learning techniques were used to explore predictive properties of these metrics. GOLD successfully predicted behavioral data as well as neural activity in response to words with varying relationships, and its predictions outperformed those of certain competing models. These results suggest that a single-mechanism account of learning word meaning from context may suffice to account for a variety of relationships between words. Further benefits of graph models of language are discussed, including their transparent record of language experience, easy interpretability, and increased psychologically plausibility over models that perform complex transformations of meaning representation.

USING A HIGH-DIMENSIONAL MODEL OF SEMANTIC SPACE TO PREDICT NEURAL ACTIVITY.

By

Alice Freeman Jackson

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2014

Advisory Committee:
Professor Donald J. Bolger, Chair
Professor Hal Daumé III
Professor Kevin Dunbar
Professor William Idsardi, Dean's Representative
Professor Meredith Rowe

Dedication

To Andrew and the future.

Acknowledgements

I was extraordinarily lucky to join the lab of DJ Bolger, a scientist of outstanding intellectual honesty, dogged persistence, and an irrepressible talent for finding time to talk. Without DJ's confident support of my work and benevolent leadership of the lab, my graduate career would have taken a very different course and this dissertation certainly would never have taken shape. I am also indebted to Tracy Riggins on this front, as her cheerful presence, advice, and generous sharing of her EEG system, lab space, and expertise made this research possible. I must thank my most patient committee members, Drs Donald J. Bolger, Hal Daume, Kevin Dunbar, William Idsardi, and Meredith Rowe, for their role in shaping and improving the present work.

I am gratified to have forged friendships with my fellow lab members, Say Young Kim, Lesley Sand, & Brandee Feola, and my fellow NACS students, particularly Katie Willis & Susan Teubner-Rhodes. I am thankful to you all for your boundless camaraderie, your celebrations when $p < .05$ and chocolatey condolences when $p = Ns$, and your bright hearts.

My education and intellectual path are as much products of my upbringing as they are of my own efforts. A lifetime of family influences is too much to express pithily, so I must distill my gratitude to its cores: I am thankful to my mother, for the Freeman traditions of curiosity and creativity; to my father, for the Jackson heritage of reverence for words and learning; and to my brother, for encouraging me in everything I've ever done.

And lastly, I have the pleasure of acknowledging my best friend and loving husband, Andrew. I am so thankful for your support, patience, encouragement, conversation, compassion, reflection, your brilliant mind, your good humor, your strength of character, and the deliberate way that you live. I am so fortunate to have you as a tremendous force of love and joy in my life.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Overview

The present study aims to develop a computational model of language that uses graphs and graph algorithms, is structured in a psychologically and/or neurologically plausible manner, and may be used to predict behavioral and neural data from language tasks. This chapter will describe how the study will progress, present relevant major theoretical issues, and summarize the research questions at hand.

1.1.1 Three major stages of the present study

The first stage is the construction of a graph-structured semantic space model, herein referred to as GOLD (Graph Of Language Distribution). GOLD will be constructed based on lexical co-occurrence within a large corpus of natural language.

The second stage is the extraction of relatedness metrics from GOLD. Metrics of word relationships will be derived from the word graph in a theoretically informed manner, such that the metrics reflect theoretical conceptions of word meaning and word relationships. This theory-driven approach will extract specific properties of the graph that correspond to theoretical constructs and use these properties to construct a variety of metrics.

The third stage of this study will comprise behavioral and neuroimaging tasks that will provide data with which to test GOLD's metrics from stage two. Specifically, the analyses of the third stage will predict (a) human ratings of word relationships and (b) neural activity in a semantic relatedness judgment task, and compare GOLD's predictive performance to that of certain existing models. Machine learning techniques will be used to discover predictive properties of the GOLD metrics; if GOLD is successful, subsequent examination may be warranted to determine if the discovered properties may further inform theory.

1.2 Major theoretical issues

1.2.1 Language representation and language models

A central question in the study of language in cognitive science is how word meaning is represented in the mind and brain. There is strong evidence that the meanings of words are learned from context (Bolger, Balass, Landen, & Perfetti, 2008a), and later reconstructed ad-hoc when meaning retrieval is necessary (Burgess & Lund, 1998; Kintsch & Mangalath, 2011). A class of computational models called 'distributional models' (discussed in Chapter 2) may be congruent with these properties of word meaning, as these models are constructed based on co-occurrence of words within a large collection of contexts, and relationships among words in the model may be later extracted. As such, these models mirror the general form of word meaning acquisition, representation, and usage as conceptualized in human language processing.

Different types of relationships between words may be considered within distributional models of language (Budanitsky & Hirst, 2005; Utsumi, 2010) and may be mathematically defined within a model (Weeds, Weir, & McCarthy, 2004). The present study will consider two different types of relationship: semantic similarity, referring to the degree of overlap in meaning features (e.g. *cat* and *feline* are highly similar, while *cat* and *blobby* are not), and associative relatedness[1], referring to co-occurrence of words in contexts (e.g. *question* and *ask* are highly associated, while *question* and *query* are not) (Budanitsky & Hirst, 2006; Kolb, 2006; Landauer & Dumais, 1997; Lund & Burgess, 1996). Distributional data may be able to capture both (Weeds & Weir, 2005), from the hypothesis that words that are similar in meaning may occur in the same role in similar contexts, while words that are associated may occur nearby. The first aim of this dissertation is to test whether GOLD can provide support for this hypothesis by calculating association from raw co-occurrence and calculating similarity from shared or patterns of connectivity between two words (Lund, Burgess, & Atchley, 1995), such that two words that are connected to the same community of words with similarly weighted connections are more similar.

It has been suggested that associative relatedness and semantic similarity are separate entities supported by separate networks of word representations, while others suggest a single mechanism of representation that can give rise to both of these relationship types (see Hutchison, 2003 for a review). However, association and similarity are not easily dissociable: words that are associated are likely to be semantically similar to some degree, and words that are semantically similar often co-occur (Deyne & Storms, 2008; Hutchison, 2003). Thus, it is difficult to argue that a particular effect arises from one relationship type or the other, as the relationships so often overlap. The present study uses a different approach: if GOLD can successfully differentiate between similarity and association, then this would suggest that the

[1] This concept is referred to by a variety of names, including semantic relatedness, association, associative relatedness, and lexical similarity. For clarity the present study will use the phrase 'associative relatedness' or 'association'.

information necessary to identify these two relationship types must be present in the single mechanism of co-occurrence.

1.2.3 The utility of computational models in brain research

A variety of computational models have been proposed that describe semantic processing of language, including acquisition of word meaning, semantic organization, and word use. These semantic models generally process a corpus of text and produce a model that represents some set of relationships among words. Some semantic models require pre-existing human analysis to specify relationships among words or concepts (e.g. WordNet, Roget's thesaurus, or Wikipedia), while others only encode those relationships that can be extracted by automated means (like distribution and co-occurrence). Specific semantic models will be reviewed in the next chapter.

Semantic models may be used for theoretical aims or for real-world applications: to judge relationships between words, like semantic distance or synonymy (Landauer, Foltz, & Laham, 1998); to make predictions of lexical items or phrases, like what word is likely to follow an existing sequence or what word a writer intended to write and instead misspelled (e.g. Islam & Inkpen, 2008); to classify input, like sorting sets of text by likely author (e.g. Burrows & Tahaghoghi, 2007); to assess the relatedness of semantic content in a student's writing to gauge how well a concept is understood (e.g. Kakkonen, Myller, Timonen, & Sutinen, 2005); and many other tasks. In light of these real-world applications, there have been concerns that these computational models are "tools" rather than valid psychological models, and while they are useful feats of engineering, they are bankrupt theoretically (Chomsky; Keynote panel, 2011). It has been argued that this is not the case (Norvig, 2011), for several reasons. Firstly, computational models are constructed based on theories of language acquisition and organization; the success of a model constructed based on a particular theory constitutes support for that theory. Secondly, computational models are typically quite parsimonious, as they are implemented manually based on a limited set of assumptions or parameters. Thirdly, computational models tend to make predictions that are well-quantified and falsifiable, which is not always the case in non-computational language models (e.g. complaints against Chomsky's theory of Universal Grammar: Piattelli-Palmarini, 1980). Lastly, a major benefit of implementing a language model in a computer is that its functioning is entirely transparent. In a computational architecture, it is known exactly what information is available to a model and what the model does with that information in order to be successful, so it is easier to draw conclusions about language processing's reliance on that information. For example, as will be discussed in subsequent chapters, many models that use only co-occurrence of words within documents have been successful at mimicking human performance on certain tasks. This success is evidence that statistical co-occurrence alone carries sufficient information to perform on these tasks. However, these models fail on other tasks (e.g. Burgess, 2000; Wiemer-Hastings, 2000), which indicates that some other information beyond co-occurrence is necessary to complete those tasks. Assessing how models achieve, or fail to achieve, their stated goals can thus further inform theory about what information the mind may use or how it may be organized.

1.2.4 Psychological and neurological plausibility of language models

The prominent distribution models such as HAL and LSA are vector space models in which words or contexts are represented as vectors in multidimensional space. Due to the vast number of words and contexts, the immensity of the vector space is necessarily reduced using an algorithm known as singular value decomposition. While highly effective as a computational tool, it is questionable whether such a process plausibly reflects a psychological process (Jones & Mewhort, 2007; Kwantes, 2005; Steyvers & Tenenbaum, 2005). It should be noted that a variety of work has explored neurally plausible implementations of complex mathematical processes, including arithmetic and more complex nonlinear computations in individual neurons (see Silver, 2010, for a review), convolution (Blouw & Eliasmith, 2003), and Fourier transforms (Velik, 2008), so it is not necessarily the case that computational models that rely on processes such as SVD can be ruled out as viable explanations of human semantic processing. However, alternatives that profess greater plausibility have been developed using episodic memory models (Kwantes, 2005), neural network models (Plaut & Booth, 2000; Rohde, Gonnerman, & Plaut, 2005), and with graph models (Collins-Thompson & Callan, 2007; Steyvers & Tenenbaum, 2005). The purported plausibility of these models arises from their congruence with cognitive theories, model assumptions, more ready interpretations of their calculations, and the types of information contained within the representations. Graph models in particular are consistent with an instance-based learning framework of word learning (Bolger et al., 2008; Daalen-kapteijns & Elshout-mohr, 2001; Fukkink, Blok, & de Glopper, 2001; Jenkins, Stein,

& Wysocki, 1984), in which episodic traces representing individual exposures to a word are accessible, but information derived from larger patterns of co-occurrence is also available. This aspect of graphs will be discussed in more detail in Chapter 2.

1.3 Research questions

The present study seeks to establish the utility of the GOLD model in predicting behavioral performance and neural activity underlying word processing. If GOLD is found to be effective, subsequent research can specify the source(s) of its predictive power. The research questions of the present study will focus on evaluating the quality of the GOLD model, and exploring what may be learned from its performance on a small suite of tasks, rather than which specific parameters of GOLD influence its performance. Each of the following three sections will introduce a finding or set of findings that GOLD is expected to replicate or outperform.

1.3.1 Can GOLD predict behavioral data?

GOLD will be used to predict human ratings of association and similarity of word pairs. GOLD is intended to capture the information necessary to judge relationships of both association and similarity from co-occurrence data. Accordingly, using theoretically informed metrics of similarity and association, GOLD is hypothesized to predict both association and similarity ratings, as well as classify words based on their relationship type. These predictions, if successful, will provide some indication the corpus is reasonable and that the methods of calculating relationships are appropriate.

1.3.2 Can GOLD predict neural data?

A specific feature of event-related potentials (ERPs) called the n400 (discussed in Chapter 2) is elicited in response to language. The n400 effect has been consistently found to be modulated by the strength of the relationship between words, such that greater relation between words in a pair produces a smaller n400 effect. Furthermore, the specific relationship types of similarity and/or association of word pairs has been shown to produce differential n400 effects (e.g. Koivisto & Revonsuo, 2001). Using similarity metrics derived from theoretical formulations of word meaning, combined with machine learning algorithms, GOLD is hypothesized to predict the size of the n400 effect elicited in response to a variety of stimuli.

1.3.3 Can GOLD's predictions outperform other models?

LSA (Landauer, Laham, & Foltz, 1997) has been used to predict amplitudes in similar electrophysiology tasks (e.g. Parviz, Johnson, Johnson, & Brock, 2011). GOLD's performance on the prediction task will be compared to LSA to determine if the GOLD is an improvement on this commonly used and broadly successful model. It is hypothesized that GOLD will outperform LSA due to GOLD's maintenance of full model dimensionality, its theory-informed similarity metrics, and its consistency with well-supported psychological theory.

Chapter 2: Literature review

This chapter aims to review relevant literature in several fields: distributional models in general, graph models in particular, event-related potentials, and machine learning. It is worth noting here that this literature review is ultimately from a perspective of what can be learned about language. Accordingly, the computer science and machine learning literatures are reviewed to the degree necessary to clarify the methods used in the present study, and are not comprehensively covered.

2.1 Distributional models

2.1.1 Introduction

The distributional hypothesis (Firth, 1957; Mcdonald & Ramscar, 2000) states that the meanings of words are related to or inferred from how words co-occur with other words in an entire corpus of contexts: if a word occurs in similar contexts as another word, then the two words should have similar meanings. The distributional hypothesis is notable in that it asserts no role of syntax, thematic organization, or even word order in inferring word meaning: the distribution of words in contexts alone is sufficient to construct their meaning. The following sections will discuss the psychological plausibility of this type of computational model, existing distributional models and their uses, and various parameters that change distributional models' utility.

2.1.2 Psychological plausibility of distribution models

Distributional models account for a wide range of behavioral findings and are strongly rooted in theory. This section will discuss two major well-supported theoretical bases of semantics that are both transparently reflected in distributional models: (1) that meaning is dynamic as well as context-constrained, and (2) that learning occurs incrementally from context.

There is plentiful evidence that the meanings of words are learned primarily from context (Fukkink et al., 2001; Swanborn & de Glopper, 1999, 2002; van Daalen-Kapteijns, Elshout-mohr, & de Glopper, 2001), that the meanings of words are fluid and dynamic (Bolger, Balass, Landen, & Perfetti, 2008; Kintsch & Mangalath, 2011) and depend heavily on context rather than formal definitions (Lawrence W Barsalou, 1987; Rogers & McClelland, 2011). Conceptually speaking, rather than looking up the meanings of words in a mental 'dictionary' when words are encountered, the meanings of words are constructed ad-hoc in a contextually-constrained manner (Burgess & Lund, 1998). Contextually-relevant meanings of words are problematic for certain other types of models, such as cognitive models of semantic knowledge that specify features or categorical organization (e.g. Mervis & Rosch, 1981), as category models can't account easily for context constraints (Rogers & McClelland, 2011). Distributional models can, as words may co-occur with other words that belong to disparate inter-connected groups that reflect different meanings.

Behavioral evidence suggests that, while acquiring meanings of novel words, learners gradually extract abstract meaning from successive exposures, while also maintaining non-abstract associations from each individual exposure (e.g. van

Daalen-kapteijns, Elshout-mohr, & de Glopper, 2001). The process of acquiring meaning gradually, through exposure to context, is formalized in the incremental learning hypothesis (Bolger et al., 2008; Fukkink et al., 2001). In a distributional framework, on exposure to a word within a context, a 'connection' between each word in the context is entered into the computational model. The unreduced distributional model thus represents the entire history of the learner's instances of exposure to language.

In human learners acquiring word meanings, a small number of exposures to a novel word leads to word knowledge that is weak and changeable (van Daalen-Kapteijns & Elshout-Mohr, 1981), and exposures to novel words in uninformative contexts leads to word knowledge that is weak or inaccurate (G. a Frishkoff, Collins-Thompson, Perfetti, & Callan, 2008; G. A. Frishkoff, Perfetti, & Collins-Thompson, 2010). In a distributional model, frequency and informativeness of exposures are both encoded: words that have been viewed infrequently or with nonspecific or generic contexts have weak connections that can be numerically overshadowed by co-occurrence with other, more informative words or by future exposures.

Furthermore, definitional meaning is not stored in a qualitatively distinct system, rather experiences of ostension are represented as an instance or contextual episode in distributional models. In such models, the core set of abstract meaning features is represented as the pattern of most frequent associates of that word. These benefits are discussed at length with respect to the HAL model (Lund & Burgess, 1996), which does not reduce the dimensionality of its representations[2] and thus maintains all of the 'memory traces' of language exposure that lead to its structure.

2.1.3 Existing distributional models and their applications

2.1.3.1 Introduction

A wide variety of computational models have been developed using distributional bases, such as LSA (Landauer & Dumais, 1997; Landauer et al., 1998), HAL (Lund & Burgess, 1996), COALS (Rohde et al., 2005), SOC-PMI (Islam & Inkpen, 2008), and many other variants. These distributional models have met with success at a variety of tasks ranging from synonymy judgment to essay grading (Kakkonen et al., 2005), indicating that the information contained just within distributions of words is sufficient to meet a surprising range of language-related goals. However, certain models that have incorporated syntactic, thematic, or other information (Kakkonen, Myller, & Sutinen, 2006; Padó & Lapata, 2006) or combined distributional models with other sources of information structure such as Wikipedia or WordNet (Agirre et al., 2009; Strube & Ponzetto, 2006) have improved on the performance of strictly distributional models in certain tasks, confirming that there is, unsurprisingly, more to language than just distribution. While distribution-only models may not reach peak performance compared to models supplemented with other information, they do possess a major advantage: models that rely only on distribution can be fully automated, and thus be reconstructed on arbitrary corpora with no additional human effort. Automation is a terrifically attractive characteristic

[2] Some variants of the HAL model do use dimensionality reduction methods, including discarding low-variance columns and multidimensional scaling algorithms (e.g. Lund, Burgess, & Atchley, 1995); it is reported that performance is equivalent between full- and reduced-dimensionality versions of the model.

when considering language, a system with a vocabulary of many hundreds of thousands of words and infinite generativity (Hauser, Chomsky, & Fitch, 2002). Accordingly, distributional models are a fruitful area of research and have been found to succeed at a wide range of tasks with real-world applications, such as grading student responses to a training program (Magliano & Graesser, 2012), synonym generation (Inkpen, 2007), scoring definitions (Collins-Thompson & Callan, 2007), authorship attribution (Burrows & Tahaghoghi, 2007), and so on.

It is worthwhile to note that computational language models relying only on co-occurrence are not intended to model the full extent of language. Some models account for other features, such as word order (e.g. Blouw & Eliasmith, 2003; Jones & Mewhort, 2007), but the majority are 'bag of words' models that discard syntactic information, and thus are incapable of making distinctions in meaning that rely on syntax, word order, or other features that are not represented in co-occurrence. Furthermore, these models are not intended to comprehend language in the sense of grounding semantic meaning in situational information (Kintsch & van Dijk, 1978). Rather, these models operate at an earlier level of comprehension (L.W. Barsalou, Santos, Simmons, & Wilson, 2008) that enables early lexical semantic processing in comprehension and word learning.

Approaches that do account for structure in language, whether syntactic or conceptual or otherwise, are profoundly valuable in the study of semantic knowledge and language, but tend to address different classes of questions than corpus-based models that rely on statistical features of language context to model relationships between units of language (Griffiths, Steyvers, & Tenenbaum, 2007).

2.1.3.2 The role of 'context' in distributional models

The distributional hypothesis asserts that the meanings of words are learned based on other words that co-occur in a context (Mcdonald & Ramscar, 2000), but it does not specify what, exactly, "context" means. It may be the case that "context" means something different in written than in spoken language. In a face-to-face conversational situation, context is not limited to the precise contents of speech and may include such factors as physical, social, and intellectual attributes of the speakers, previous topics discussed by the speakers, prosody, and so on. It may be the case that all of these contextual cues are relevant in interpreting or constructing (Kintsch & Mangalath, 2011) the meaning of an utterance. However, in developing semantic space models, context is assumed to be limited to the words present in the current text.

In semantic space models, words count as co-occurring with a target word if they fall within some "window" of words around the target word in a text. Models may use several sizes of windows: some use 'document' as the smallest organizational unit, and link every word in a document to every other word (e.g. LSA: Landauer, Foltz, & Laham, 1998) others use some smaller value (e.g. ten words before and after the target word: Lund & Burgess, 1996). These models typically slide the window over the entire document, counting co-occurrence to the target word in the center of each window until the end of the document is reached. The role of window size in model performance has been assessed (e.g. Bullinaria & Levy, 2012) with the general finding that increasing window size produces worse performance. However, this analysis was carried out using models that collapse the dimensionality of the represented corpus; it is unclear if this finding will apply to models that preserve dimensionality (dimensionality is discussed below).

Naturalistic texts provide additional meaningful units of organization beyond the ‘document’, namely the sentence and the paragraph. There is evidence that these organizational units are reflected to some degree in a reader’s processing of the text (e.g. Goldman, Hogaboam, Bell, & Perfetti, 1980; Ledoux, Camblin, Swaab, & Gordon, 2006; Shanahan, Kamil, & Tobin, 1982).

2.1.3.3 The role of corpus size and selection in distributional models

Selecting an insufficiently large corpus carries two risks: first, that a word may not be represented at all in the corpus, and second, that all of the senses of the word may not be represented in the corpus. What constitutes a “large” corpus has varied dramatically over the years: versions of LSA by 1997 used “very large numbers of words” in the range of 20-70k (Landauer et al., 1997); early HAL models (Lund & Burgess, 1996) used 160 million words from USENET; HiDex, a later HAL-type model, used a one billion word corpus from USENET (Shaoul & Westbury, 2010), in part because a 160 million word subset did not include every word from their 50,000-word lexicon. If a corpus contains no instances of a word, then clearly that word is not represented and cannot be processed using the resulting model; if a corpus contains very few instances of a word, it is unlikely that those instances span all possible senses in which a word may be used. As English is rife with polysemy (84% of words examined in Rodd, Gaskell, & Marslen-Wilson, 2004), a small corpus might be expected to exclude alternate meanings or uses of a huge number of words. Hence, larger corpora should be more likely to capture the variance with which words are used – not only increasing range of associations, but also allowing the model to encounter words with multiple meanings in many different contexts.

A small corpus also risks insufficient representation of domain-specific terms. For example, while CPU and RAM have specific meanings whose differences are vital to the workings of computers, LSA-type models judge the two terms to be highly similar, in some cases maximally similar (Wiemer-Hastings, 2000). Both occur in a specific domain – a computer's hardware – and either the limited corpus or the dimensionality reduction eliminated the fine distinctions between the two terms.

It may be valuable from a perspective of ecological validity to construct models that mimic human experience, but many existing models use corpus sizes that do not reflect the size or range of realistic language input to a developing human. It is difficult to estimate how many words a person hears and reads over the course of a lifetime, but a lower bound may be estimated using the Human Speechome Project[3], which recorded the in-home audiovisual environment of a child from infancy to age three. A subset of the recordings has been transcribed, yielding a set of 7 million (total, non-unique) words to which the child was exposed[4]. Considering that not all of the records had been transcribed, and that the entire dataset represents only three years of exposure to speech and minimal exposure to written text, it seems safe to place a (very) conservative lower bound of exposure to language at 7 million words. A more appropriate lower bound estimate would scale this figure by age, such that an 18-year-old would have heard six times more than a 3-year-old, leading to a figure of

[3] http://www.media.mit.edu/cogmac/projects/hsp.html
[4] http://www.ted.com/talks/deb_roy_the_birth_of_a_word.html

42 million words; this figure accounts only for spoken, and not written, words. In either case, theoretically, corpora sizes on the order of millions would be more ecologically valid than smaller corpora.

From a data-driven standpoint, there is strong evidence that vastly increasing the size of a corpus can lead to increased success using a distributional model (e.g. Chelba, Bikel, Shugrina, Nguyen, & Kumar, 2012; Dean et al., 2012). Some studies have found diminishing returns beyond some threshold size (90 million words, in Bullinaria & Levy, 2007), while some have found unbounded benefits at larger corpus sizes (2 billion words, in Bullinaria & Levy, 2012). The utility of larger corpora may also depend on the measure in question: there is evidence that simply increasing the size of the input corpora can dramatically improve performance at certain automated tasks, especially if the corpus comprises unlabeled data (Dumais, Banko, Brill, Lin, & Ng, 2002; Recchia & Jones, 2009). Whether or not more data will improve performance in the present model is a directly testable question, as the data are collected and then stored in units of documents, and thus document sets of varying size may be tested in the same way, and their performance compared. Addressing this question is beyond the scope of the present study, but may be addressed in future work.

2.1.3.4 Manually annotated taxonomies

A number of studies have examined the utility of word relationships that have been manually defined or organized, such as dictionaries, thesauruses, and knowledgebases like Wikipedia or WordNet (Miller, 1995). Budanitsky & Hirst (2005) reviewed a variety of human-organized knowledge bases (e.g. Roget's

Thesaurus, WordNet (Miller, 1995), MeSH[5]) and compared the performance of various similarity metrics trained on WordNet's human-annotated data; a variety of other works have used knowledgebases entirely, or in combination with language distributions, to complete language tasks (e.g. Agirre et al., 2009; Gabrilovich & Markovitch, 2007; Jarmasz, 2003; Li, Sun, & Datta, 2011; Mihalcea, Corley, & Strapparava, 2005; Strube & Ponzetto, 2006). These models typically perform very well, which is one of many arguments to be made in support of manually constructed knowledgebases. However, human-annotated models suffer from the general limitations of (a) the enormous amount of time required to annotate or organize the data, (b) that only data that has been preprocessed in this resource-intensive manner can be used by the model, and (c) the assumption that the structure of meaning in language is both static and predefined. These models require a correct, precise taxonomy of terms and concepts, which depend on extensive and accurate human effort. In contrast, an automated system lacks the additional information that is provided by human judgment, but is cheaper, faster, and much less limited in scope.

Another major drawback of human-annotated corpora is that the model is 'frozen' in the historical period in which the model was made, and cannot incorporate novel uses of language without massive human effort. It is an often-lamented reality that language is continually evolving (e.g. Dorogovtsev & Mendes, 2001; Scheel, 1998). A human-annotated model generally only captures a 'snapshot' of a language, while an automated processor can track evolving language use in a community on a much shorter timescale than the years it takes to complete a project on the scale of WordNet.

[5] http://www.ncbi.nlm.nih.gov/mesh

2.1.3.5 Model dimensionality

Natural language is vast. The OED contains 600,000 unique words[6], while the Google Books project has estimated that English contains over a million unique words (Michel et al., 2011). Given the enormous size of the vocabulary, much less the possible combinations of multiple words into phrases, maintaining the full dimensionality of a language-derived space has traditionally been difficult. Some models maintain most of the dimensionality of the semantic space, notably the HAL model (Lund & Burgess, 1996), which performs well at extracting both similarity and association, as well as additional tasks such as categorization. Many existing models do collapse across dimensions using procedures like singular value decomposition (in LSA; Landauer et al., 1997) or various approaches that discard dimensions based on their variance (Lund, Burgess, & Atchley, 1995) to yield a much more manageable computational space, however these reduced dimensions (a) do not map directly to concepts or words, and (b) necessarily minimize the salience of less dominant meanings of words. Some have argued that the real dimensionality of the human semantic space is very small (Lowe, 2000), and thus that dimensionality reduction accurately reflects human semantic processing. However, compressional/reduction methods like SVD have been found to distinguish poorly among near-synonyms (Wang & Hirst, 2010) or multiple meanings of words (Lee, Baker, Song, & Wetherbe, 2010). These findings indicate that, from a data-driven perspective, higher-dimensional representations may be necessary for at least some tasks of language use.

2.1.3.6 Word frequencies

Lastly, this method of model construction also produces word frequency counts. Word frequencies are strong predictors of reaction time in a wide variety of

[6] http://public.oed.com/about/

reading tasks; accordingly, the accuracy of the model of language from which word frequencies are derived is critical (Burgess & Livesay, 1998). The word frequency counts expected from this internet-based corpus may more accurately reflect the language experience of participants than many existing word frequency databases. Consider that the word *pizza* has the same frequency as *scrutiny* in the American National Corpus[7] and *advocate* in the BYU Contemporary American corpus[8], and it doesn't even appear in the Brown corpus (Wilson, 1988). Given that the target population of most university studies is the infamous college sophomore, a corpus based on language generated by many users (many of whom are from a college demographic) may be a better fit for experimental uses.

It has been found (Burgess & Livesay, 1998) that a larger and more recent set of frequencies (from the HAL corpus: Lund & Burgess, 1996) more strongly predicted medium-to-low frequency words than the Brown corpus. High-frequency words in a language are less likely to change or be replaced by new words over time (Pagel, Atkinson, & Meade, 2007), which may explain the older Brown corpus predicted reaction times to high frequency words as well as the newer corpus. Accordingly, a corpus that reflects realistic, conversational word frequencies – and can be updated automatically to reflect changing language – may be ideally suited to experimental use.

[7] http://www.anc.org/frequency.html
[8] http://corpus.byu.edu/coca/

2.2 Graph models

2.2.1 Introduction

The majority of the models discussed in the preceding section are vector space models in which words or sets of words are represented as vectors in a dimension-reduced space. Far fewer researchers have used a graph theory approach to constructing models based on the distributional hypothesis, though these models are rapidly gaining traction (Radev & Mihalcea, 2008). This section will introduce graphs and discuss some graph models that have met with success in previous research.

Graphs are methods of representing data and relationships among data using ‘nodes’ and ‘edges’ or ‘connections’. Connections between nodes have an associated number referred to as ‘weight’. In the case of a graph model of language, each node may represent a word, a document, and the weight of a connection between two nodes may represent proximity or frequency of co-occurrence. A possible benefit of graph models of language is that the data are not necessarily collapsed or reduced, though reduction is possible. Instead of singular value decomposition (SVD) or similar algorithms needed for high dimensionality models, reduction of complexity in graphs may be executed using clustering, by collapsing clusters of nodes into supernodes that could be described as latent concepts, by directly collapsing synonyms, or by pruning of nodes or connections based on weights, frequencies, or other properties.

2.2.2 Existing graph models

Graph models that have been used in the literature have varied widely in the target tasks and algorithms employed. Previous research has addressed the task of identifying category exemplars using an algorithm that considered each new exemplar candidate's connectivity to previously identified exemplars (Widdows & Dorow, 2002); gauged document similarity using a type of sub-graph comparison that compared the entirety of the documents rather than considering individual terms (Tsang & Stevenson, 2010); and identified 'communities' corresponding to word senses using clique analysis, an algorithm commonly applied to social networks (Palla, Derényi, Farkas, & Vicsek, 2005). The MESA model (Collins-Thompson & Callan, 2007) used random walk Markov chains through a graph whose connections represented several different types of word relationships to judge the quality of word definitions, while Huges and Ramage (2007) used random walk Markov chains on graphs based on WordNet relationships to judge semantic similarity of word pairs. The consistent feature of these studies is that each study exploits graph-specific properties of the model and graph analysis algorithms to address their chosen tasks.

The combination of graph models with machine learning approaches has also been successful at various language tasks. Machine learning algorithms may be used to find patterns in existing data, and use those patterns to predict characteristics of new data. This approach may be particularly useful when the model produces or contains a great deal of information, but is not clear on precisely how that information should be combined or reduced to a final prediction. Minkov and Cohen (2008) combined a graph theoretic approach with machine learning techniques to learn a similarity metric with a graph walk algorithm. Silva and Amancio (2013) used specific types of graph traversal with machine learning classifiers to perform word sense disambiguation. The combination of graph theory and machine learning may be fruitful, as graph analysis algorithms may extract information from the word graph that can then be used as inputs to the machine learning algorithm.

2.2.3 Psychological and/or neurological plausibility of graph models

Graph models[9] provide certain additional relevance to the psychological study of language, largely stemming from the fact that dimensionality of the model is not reduced in any transformative manner. While low-frequency words or low-weight connections may be deleted from a graph model in order to reduce its computational burden, these deletions don't impact any other words or connections. Each node still represents a word and each connection still represents first-order co-occurrence. In contrast, the matrix reduction used in LSA takes a semantic space with many thousands of dimensions and reduces it to a few hundred dimensions, such that vectors within the resulting space do not correspond directly to any specific concepts (hence the 'latent' meaning in 'latent semantic analysis').

A major benefit of full graphs of co-occurrence, rather than reduced vector spaces, is that the full graph allows statistical properties of language to accrue from the episodic traces that are reflected in connection weights (Kwantes, 2005; Steyvers & Tenenbaum, 2005), grounding the graph in the episodic-trace models of memory (Hintzman, 1984; Howard, Addis, Jing, & Kahana, 2005; Kwantes, 2005). Thus, maintaining full dimensionality in a graph model doesn't eliminate information as singular value decomposition does. Instead, it records the history of language exposure in very clear way and allows for easier interpretation of model output because nodes and edges reflect specific words and co-occurrence, rather than latent

[9] Graphs can be represented as matrices, and thus information within a graph may still be described as vectors.

meaning (Audet & Burgess, 1999; Burgess & Lund, 1997; Lund & Burgess, 1996). The ultimate output of the graph model – in this case, judgments of similarity and association – is thus extracted from the accumulation of contexts that contain the target words. This is a mechanism that is consistent with theories of word learning, particularly the instance-based learning framework (Bolger et al., 2008), that assert that the meanings of words are learned from features that are consistently present in discourse or other contexts.

2.3 Event-related potentials

2.3.1 Introduction

The preceding sections reviewed research in language models. The success of the language model in the present study will be quantified by its ability to predict neural activity as measured by event-related potentials (ERPs). Accordingly, the following section will introduce ERPs and discuss their utility in studying language processes.

ERPs are small segments of electroencephalograph (EEG) recordings that are time-locked to the onset of stimuli and averaged over many trials to produce an averaged waveform. Averaging many trials allows a very small event-related signal to be extracted from the background noise of brain activity. Various features, referred to as *components*, of the time-locked waveform have been identified as reflecting particular language-related processes or experimental manipulations (Kaan, 2007; Osterhout, Kim, & Kuperberg, 2006). Several of these ERP components have been used as tools to examine various aspects of on-line processes involved in reading, among them the n400 (discussed below).

There are many benefits of collecting ERP data in addition to behavioral data, notably their sensitivity. ERPs are generally considered to be more sensitive than behavioral output, for several reasons. Firstly, ERP data are high-dimensional: 64 or 128 channels and generally around a thousand timepoints per trial. While a task with a yes/no response generally only examines variance on the two metrics of reaction time and accuracy of a decision output, ERP can allow the examination of latent activity that is collapsed into the single instance of behavioral output. In the present study regarding word knowledge, if variability in knowledge or representation of a word is not large enough to produce different behavioral output, or if the variability is on a dimension that doesn't directly alter behavioral output on a particular task, then the variability may not be reflected in behavior. ERPs provide a sensitive measure that is often able to measure such latent variability in cognitive processes.

2.3.2 The n400 component

The n400 is a negative deflection in the EEG signal that occurs roughly 400ms after stimulus onset. This component, extensively reviewed elsewhere (Kutas & Federmeier, 2011) is commonly used as an index of semantic knowledge and integration of semantic knowledge into existing contexts. Of particular importance is that the degree of relationship between a predicted target and the actual target has been found to modulate the amplitude of the n400 (e.g. Federmeier & Kutas, 1999), and that similarity between word pairs in priming tasks shows a similar, though sometimes attenuated, effect (Perfetti, Wlotko, & Hart, 2005). Koivisto and Revonsuo (2001) found that both semantic similarity and relatedness affect n400 amplitude, but noted that related words elicited a longer-lasting n400 priming effect than the similar words. These properties make the n400 an ideal tool for investigating language processes and how word meanings are represented or manipulated in the brain.

2.4 Machine learning

2.4.1 Introduction

Machine learning (ML) uses ‘features’, or predictors, and ‘examples’, or instances of data from which to learn or to predict. In the present study, the output of the GOLD model will make up the features and grand average ERPs will make up the examples. Many features will be used as inputs to the ML algorithms because the literature informs no specific pre-existing hypotheses about which types of similarity calculation and/or normalization are most appropriate. It may be valuable to use feature selection, in which predictions are made using only a subset of features that have been identified as being more informative than others, particularly because many of the GOLD features will be correlated. Feature reduction often leads to better performance, except in the case where certain features predict a subset of the problem space that other features do not predict (Hall, 1999). Additionally, variables that are correlated can still add information, as long as they are not perfectly correlated (Guyon & Elisseeff, 2003). Accordingly, the present model will rely on the full set of features from GOLD as well as exploring model performance with reduced sets of features.

2.4.2 Types of algorithms

No a priori hypotheses regarding ML algorithms, so naïve implementations of several different algorithms were tested, including support vector machines, neural networks, random forests, and k-nearest-neighbors. Each of these algorithms is briefly introduced below.

Support vector machines (SVMs) and support vector regressors (SVRs) can identify patterns in data that are complexly related by mapping the data into a new space in which they are more simply related. Furthermore, SVMs/SVRs aim to optimize these transforms such that the space between the classes of examples is as wide as possible, which allows for better generalization. These methods are robust in the face of noisy and/or sparse, high-dimensional, and have been used with success in brain research (Lotte, Congedo, Lécuyer, Lamarche, & Arnaldi, 2007) and a variety of other fields.

Neural networks (Cheng & Titterington, 1994; Hopfield, 1982) are based on a very simplified model of neurons, typically modeled as layers of ‘neurons’: an input layer, one or more hidden layers, and an output layer (the present study uses multilayer perceptrons with a single hidden layer). The input layer takes in the stimuli, passes them on to the hidden layer, and the hidden layer outputs to the output layer which corresponds directly or indirectly to the network’s decision. All of the connections between neurons in each layer are weighted, and those weights altered such that the pattern of weights in the network can represent transformations from input to output. Neural networks have been applied to a variety of fields including language research (Bengio, Ducharme, Vincent, & Jauvin, 2003)

The random forest algorithm (Breiman, 2001) trains many decision trees that are initialized with random weights. Instead of relying on a single decision tree's prediction, it averages over the predictions of all of the trees in the forest, to produce an output that is more robust against noise and vagaries of random weight assignment. Random forests have met with success in language modeling (Xu & Jelinek, 2004).

The *k*-nearest-neighbors algorithm considers the *k* training examples that are nearest in the feature space to a test example, and assigns the average value (for regression) or most common class (for classification) of the neighbors as the prediction of the test example. This is a fairly simple approach, and considers only the immediate feature space, but achieves high performance on a variety of measures (e.g. Weinberger, Blitzer, & Saul, 2009).

2.4.3 Psychological/neurological plausibility

In keeping with the theme of psychological/neurological plausibility, it seemed appropriate to restrict GOLD's learners to algorithms that are plausibly implementable in a brain. However, what exactly constitutes a psychologically or neurologically plausible mechanism is not clear. Logically speaking, it is the case a neural network of suitable size with one or more hidden layers is capable of performing arbitrarily complex mathematical operations (Hornik, Stinchcombe, & White, 1989); if the brain can operate as the mathematically modeled neural networks do, then it is not obvious that an algorithm like SVM, or even SVD, could not be occurring in the brain. Empirically speaking, realistic models of neurons have found success at modeling a variety of algorithms, including fast Fourier transforms (Velik,

2008) and convolution (Blouw & Eliasmith, 2003). Accordingly, it seems inappropriate to rule out a particular algorithm based on its implausibility, and so all of the aforementioned ML algorithms will be used and discussed.

2.5 Summary

This chapter reviewed relevant literature in language acquisition and representation (the distributional hypothesis), semantic space models, graph models, language-related ERPs, and the basics of machine learning. This past work leads to the general hypothesis that a graph model of distributional data may give rise to similarity measures that can predict behavior as well as neural activity measured via ERP. The next chapter discusses the construction of such a model.

Chapter 3: Methods

This section will describe the construction of the GOLD model, the LSA model, and the machine learning techniques that will be used to predict behavioral and brain data.

3.1 GOLD model

3.1.1 Introduction

The present study will construct a graph-structured model (GOLD) of English based on the distributional hypothesis discussed in the previous chapter. The ultimate goal of GOLD in the present study is to measure similarity of two sets of words by representing their meanings through their relationships to other words.

GOLD will not reduce its complexity to a small set of dimensions as in LSA (Landauer et al., 1997) and many other vector space models. Instead, GOLD will take the form of a graph in which each node represents a word and the weights associated with connections between nodes will represent relative frequency and proximity of co-occurrence. The weakest connections between nodes and/or the most infrequent words may be removed from the graph in the interest of reducing necessary computations, and connection values may be normalized, but no further transformations will be applied. Maintaining, rather than reducing, the dimensionality of the data is intended to allow the finest possible comparisons between words by not eliminating any information about their connectivity.

3.1.2 Corpus

In an attempt to capture modern language usage, we collected a corpus from comments on the forum website Reddit (www.reddit.com), which is one of the most frequently visited websites on the internet (www.alexa.com). The benefits of using a Reddit comment corpus include naturalistic language use, a wide range of authors, a broad array of topics under discussion, and a vast pool of data. Posts in the most popular subsections of Reddit (enumerated at http://subreddits.org/) were queried roughly daily from October 2012 through February 2013, and threads containing more than 100 comments were collected. Comments were parsed at the 'document' level, which consisted of the entire comment thread; the 'paragraph' level, which took <p> and
 tags as paragraph breaks; and the 'sentence' level, which used sentence-final punctuation such as periods and exclamation points as delimiters in addition to the paragraph breaks. The GOLD model was constructed based on the paragraph level data, as a compromise between the computational complexity of full-document processing and the limited span of the sentence-level data. A total of 19,646 comment threads were collected, totaling 4,342,302 paragraphs, 97,976,253 tokens (word instances), with 431,822 types (unique words).

3.1.3 Preprocessing

The corpus was stripped of several classes of letterstrings. Stop words (closed-class words such as *it, the, and;* using NLTK's English 127-word stoplist; Bird, Loper, & Klein, 2009) were removed, on the premise that removal of stop words does not impact the output of the network but does dramatically decrease the computational load of network construction and analysis (Bullinaria & Levy, 2012).

This removed 50,064,361 tokens, more than half of the corpus. Unique strings that did not occur in a large set of words combined from NLTK's word lists (size 755,110) and NLTK's package of WordNet (size 10,771,928) were removed on the premise that these words are not common terms in the language. This step eliminated letterstrings such as *fooooood, hasbut,* and *qxt,* and protowords such as *facepalm, derp,* and *awesomesauce*. A surprising 362,202 types were removed in this step, for two reasons. First, retaining only words that occur in wordlists is overly conservative, as many legitimate words were not present in the wordlists (such as *minnesota* and *minecraft*). Second, the internet is rife with creative misspellings, and these strings are more likely to be unique than correct spellings – for example, *someone* may occur with a high frequency but only count as a single unique type, while *sumone, someon, somoen, summone,* etc., will each count as a separate, unique type. Despite the huge number of types removed in this step, these types accounted for only 2,112,017 tokens, or ~2.15% of the corpus. Lastly, strings that occurred only once in the entire corpus (10,592 tokens, such as *osseous* and *monomorphism)* were removed on the premise that very low frequency words will be connected to a very small set of co-occurring words and thus cannot contribute much to the network processing or to psychological meaning.

A final list of 58,901 types remained after cleaning, composing a corpus of 45,799,875 tokens.

3.1.4 Constructing the graph

Co-occurrence of words within the cleaned corpus was calculated by examining each paragraph in turn, pairing every word in the paragraph with every other word, and incrementing the weight of the connection for each word pair by 1. Paragraphs of length=1 (e.g. "cuuuuuuuuuute" and, mysteriously, “onychomycosis”) were ignored. The total collection of word pairs and connection weights were fed into graph database software (Neo4j version 1.8.2; Eifrem, 2009) to construct the graph. A total of 58,901 unique words (nodes) and 54,399,032 weighted relationships among those words (edges) were included in the GOLD model. The graph possesses expected properties of a large-scale language network (Steyvers & Tenenbaum, 2005), such as a degree distribution following Zipf’s law and small-world structure.

On the advice of Bullinaria and Levy (2007, 2012), the network was reconstructed using a window of size=1, such that words were only connected to words that occurred immediately adjacent in the cleaned paragraphs. This network included 58,901 nodes and 10,603,851 weighted edges, and is hereafter referred to as ‘smallGOLD’.

Figures 1 and 2 display the immediate neighbors of two pairs of words in smallGOLD: *grumpy-cat* in Figure 1, and *sushi-octopus* in Figure 2. Figure 1 is too dense to discern much about individual connections, but in Figure 2, edges’ thickness and color reflect their weight. The effect of frequency is very apparent in Figure 1, as *grumpy* occurs 754 times in the corpus, while *cat* occurs 17,551 times; accordingly, the size of the *cat* associate cloud dwarfs that of the *grumpy* associate cloud. Figure 2 displays a pair that is much closer in frequency: *sushi* occurs 938 times in the corpus, while *octopus* occurs 512 times. It is worth noting that the higher frequency words are more likely to be in the overlap set (those nodes that are connected to both words of the word pair) merely as a result of frequency.

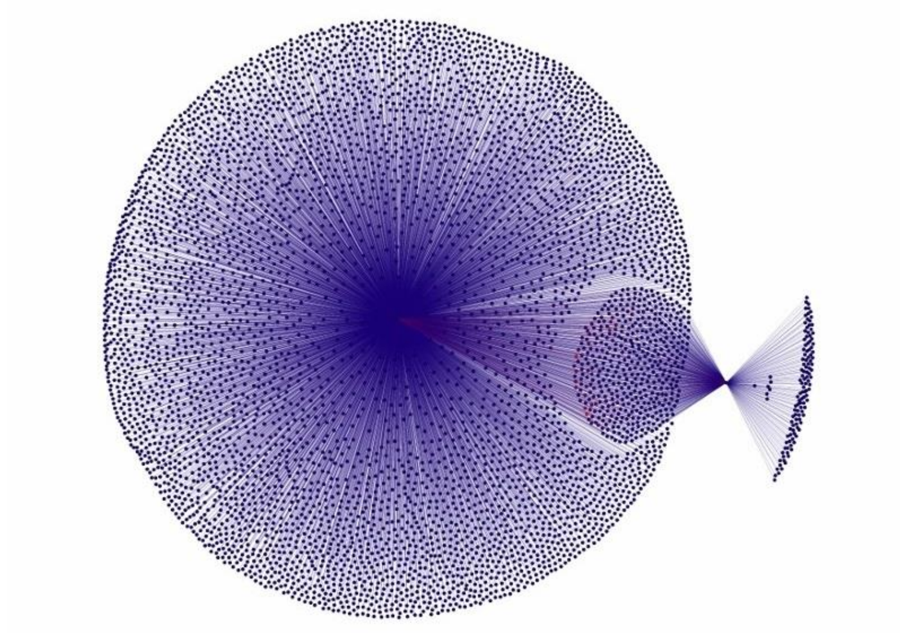

Figure 1. First-order associates of grumpy-cat. Connectivity between associates is not displayed. The large cloud of nodes are the associates of *cat* that are not also connected to *grumpy*; the small cloud of nodes are the associates of *grumpy* that are not also connected to *cat*; and the round blob between them is the set of nodes that is connected to both *grumpy* and *cat.* Figure produced using Force Atlas and Yifan-Hu layout algorithms in Gephi (Bastian, Heymann, & Jacomy, 2009).

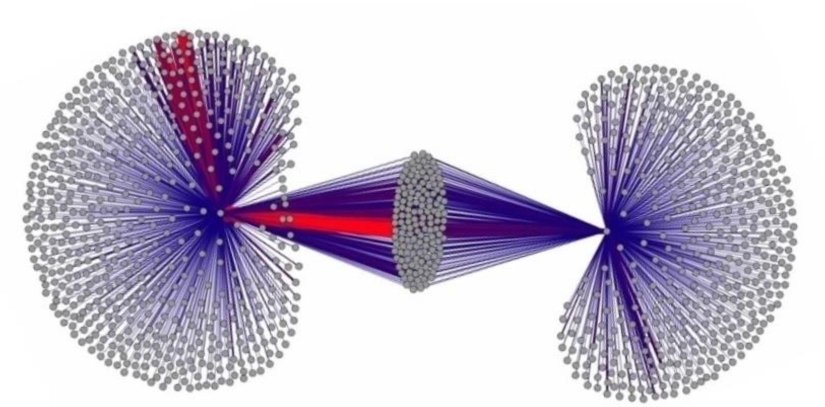

Figure 2. First-order associates of sushi-octopus. Connectivity between associates is not displayed. This subgraph is small enough to display weight information as well; weight of connections is depicted by color (red=large weights) as well as thickness. Figure produced using Force Atlas and Yifan-Hu layout algorithms in Gephi (Bastian et al., 2009).

3.1.5 Normalization

Theoretically, high-frequency words carry less information or specificity of meaning than low-frequency words (Finn, 1977; Schatz & Baldwin, 1986). That is, terms with high specificity are used more rarely because their specificity is applicable more rarely (e.g. the concept denoted by *antidisestablishmentarianism* isn't relevant often in daily life). In contrast, more frequent words tend to be far less specific and are more likely to be polysemous (e.g. *run*). In a co-occurrence model, high-frequency words are connected heavily and widely merely as a product of their frequency, rather than necessarily reflecting meaningful relationships. Accordingly, these abundant, heavy weights must be normalized to remove this undue influence of frequency. Any applied normalization method must account for frequencies of the words at both ends of an edge; several standard methods, such as pointwise mutual information (PMI) and association strength (Eck & Waltman, 2009) already do this, while other methods that only normalize node properties, such as inverse document frequency (IDF), may be altered to suit a two-word relationship. The theoretical underpinnings of graph models of language are clear that weights should be normalized, but are not clear on the best manner of normalizing weights. Accordingly, we used 15 different normalization techniques that rely on combinations of raw frequency, document frequency, IDF, and log transforms of these frequencies.

3.1.6 Similarity and association metrics

There is evidence (e.g. Weeds & Weir, 2005) that examination of different types of information within a model framework can identify different types of relationships such as similarity and association. From a theory-driven perspective, the structure of a word graph may be able to directly capture both types of relationships. Semantic similarity between two items may be reflected in second-order connections, or the intersection between their connections (i.e. are both words connected to the same set of other words?). Association may be captured in first-order connections, or the connection between the two items themselves (are the words connected to each other? If so, how strongly?). These proposed patterns derive from the distributional hypothesis, for the following reasons. Similarity would be represented in second-order connections because two words that connect to the same neighborhood of words may take the same role (e.g. *the hot cup of coffee* and *the warm cup of coffee*); similarity would not be captured in first-order connections because natural language doesn't generally provide that kind of redundancy (e.g. *the hot and warm coffee).* Association would be represented in first-order connections because those would co-occur directly together, as *coffee* and *hot* would be associated in the previous example, as would *coffee* and *warm*.

From a data-driven perspective, it may be beneficial to view the model as containing useful information of some kind, but remain agnostic as to the exact form of that information. Machine learning techniques will be used to discover and describe, rather than proscribe, what properties of the word graph may be useful in representing different relationships between words. However, theory will inform the properties that are extracted from the graph to be input to the machine learning algorithms. The use of both theory and data to inform model metrics will be useful on several levels. The theory-driven approach is more clearly informed and psychologically valid; the data-driven approach may yield a metric that is more difficult to interpret psychologically, but will produce more accurate predictions. If this is the case, the metrics may be examined more closely to determine what sort of information in the graph it is relying on to produce better predictions, which may in turn inform theory. In this way, if existing theory is incomplete in explaining how relationships are encoded in distributional data, the data-driven method may be used to discover additional factors that might make theory more complete.

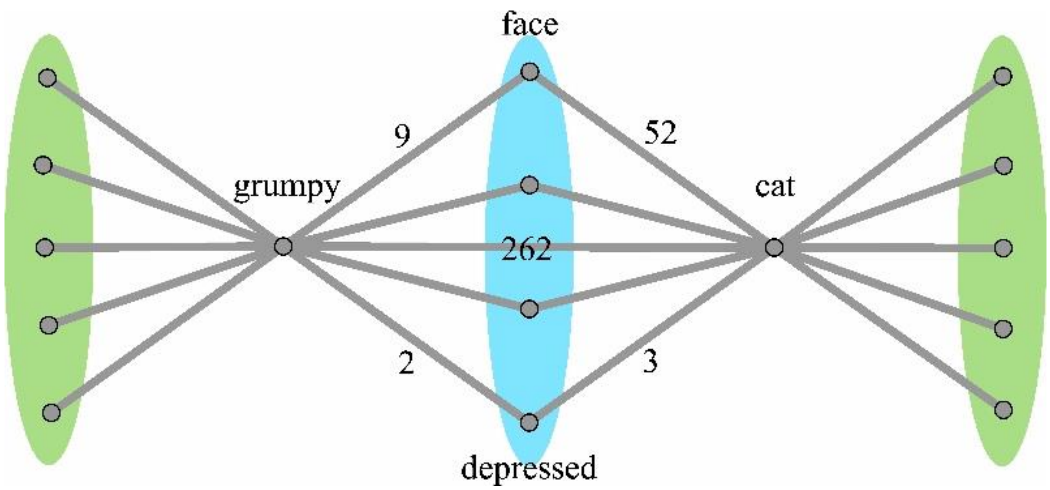

Figure 3. A simplified graph of grumpy-cat. Overlap nodes are shown on a blue background and nonoverlap nodes are shown on a green background.

Ideal metrics for assessing relatedness between words in the GOLD model should (a) reflect psycholinguistic theories, (b) preferably be limited to a set range of values, such as LSA's -1 to 1, for easy comparison, and (c) differentially consider nodes that are connected to both words in a word pair as well as words that were uniquely connected to each word, as both first- and second-order co-occurrences putatively contribute to relatedness differentially. Figure 3 presents a very small subset of the associates of *grumpy-cat* to illustrate the overlap and nonoverlap nodes.

Association was theorized to be reflected in the direct connection between the two words in a word pair, which reflects the episodic history of how often the two words co-occur. This metric has no upper bound, and a minimum of 0 indicating no relationship. This metric was calculated by extracting the raw weight of the connection between the two words and normalizing it by the normalization methods in Table 1. An additional metric was determined by calculating PMI as follows, where *w* is the weight between the two words in the word pair, $w_1 df$ is the document frequency of word 1, and n_{docs} is the total number of documents in the corpus:

$$PMI = \log_{10}\left(\frac{w * n_{docs}}{w_1 df * w_2 df}\right)$$

Additionally, 15 methods of normalizing the connection weights were used (see Table 7 in Appendix A for normalization methods). All permutations of these association algorithms and normalization methods were calculated from the graph, for a total of 30 association metrics (15 normalization methods x 2 association calculation methods).

Semantic similarity goes beyond the simple co-occurrence between to words and is theoretically reflected in shared or overlapping patterns of connectivity for two words (Lund, Burgess, & Atchley, 1995), such that two words that are connected to the same community of words with similarly weighted connections are more similar. In essence, the graded nature of similarity (e.g. Collins & Loftus, 1975) might be represented by some combination of the overlapping relative to non-overlapping patterns of connections and the fundamental weighting of those connections. This general conception of similarity is akin to Lin's universal similarity measure (Lin 1998b, as reviewed in Budanitsky & Hirst, 2005), although with a definition of overlap that arises from connectivity rather than information directly.

This theoretical conception does not prescribe the exact calculation of the metric, so in order to determine the optimal metric for detecting similarity versus association in GOLD, we tested 5 different algorithms (see Appendix A for calculation details). All permutations of the similarity algorithms and normalization methods were calculated from the graph, for a total of 75 similarity metrics (15 normalization methods x 5 similarity calculation methods). These metrics are redundant to some degree; however, because one of the primary goals of the present study was to establish if the information necessary to classify stimuli is present in the graph, the full set of metrics was input into the neural network classifiers. Additionally, eliminating metrics based on performance on this stimulus set may provide an inaccurate view of which metrics are necessary or most predictive, because this stimulus set is not designed to span the full space of relationships (e.g. there may be many synonyms and few antonyms in the stimulus set).

3.2 Latent semantic analysis (LSA)

Latent Semantic Analysis (LSA) is a vector-space model commonly used in language research to gauge word relationships and is often considered the gold standard for performance of a range of measures. Accordingly, LSA was used here as a comparison model. LSA was constructed on the corpus described above using gensim (Rehurek & Sojka, 2004). The same preprocessing steps were applied to the corpus and the model was constructed with 300 dimensions, as has been determined to be optimal for LSA model creation for a variety of tasks (Landauer, Laham & Foltz, 1997).

3.3 Machine learning

In both Experiment 1 and Experiment 2, model predictions were quantified using the Orange machine learning software suite (Demsar et al., 2013). Classifiers were trained for tasks that required sorting stimuli into discrete groups and regressors were trained for tasks that required predicting continuous values, using the algorithms described in section 2.4.2.

3.4 Summary

Chapter 3 described the construction of the GOLD model and an LSA model. These models will be used to predict rating data in Experiment 1 in Chapter 4, and neural activity in Experiment 2 in Chapter 5.

Chapter 4: Experiment 1 (behavioral data)

Assessing relationships between words by asking participants to make rating judgments is a commonly used method that dates to at least the 1960's, with Rubenstein and Goodenough's (1965) experimental validation of contemporary theories of conceptual similarity. Rated word pairs of this nature are often used as standards of comparison for computational models of language (Budanitsky & Hirst, 2006; Kintsch & Mangalath, 2011) as they are thought to reflect theoretical accounts of semantic knowledge as well as empirical human judgment.

4.1 Stimuli

4.1.1 For human subjects in Experiment 1a and Experiment 2

The stimulus set was limited to 350-400 word pairs based on the duration of each trial (~4s, plus ITI) and the tolerance of participants to lengthy sessions. Word pairs were drawn from existing studies (Chiarello, Burgess, & Richards, 1990; Thompson-schill, Kurtz, & Gabrieli, 1998; and Miller & Charles, 1991 and Rubenstein & Goodenough, 1965 as cited in Budanitsky & Hirst), and then additional word pairs were generated from the Reddit corpus. First, the lexicon of the cleaned Reddit corpus was reduced to words with frequency > 100 and length > a2. Words appearing in a taboo word list (words referring to racial slurs, explicit violence, etc.) were removed. Then, the following procedure attempted to produce a stimulus set from these words that spanned the relatedness space. Ten thousand words were randomly selected from the reduced word list. These 10,000 words were randomly paired several times and sorted into bins based on their LSA cosines[10]. Two hundred word pairs from each of the 15 LSA bins were randomly selected, and those pairs were further whittled down by removing word pairs containing a word with multiple meanings.

Because word frequency can influence behavior and neural activity, an attempt was made to balance words pairs in each bin on frequency, such that the average frequencies of words in each bin were equivalent, by removing word pairs with extreme frequency values (both high and low). However, this attempt was not entirely successful, because higher frequency words tend to have higher cosines with other words of high or medium-high frequency. It was more likely that word pairs that are unrelated according to LSA are also lower frequency, so the most unrelated bins have a slightly lower average frequency (see Appendix D).

Many words were duplicated between the word pairs drawn from other studies and the randomly generated pairs. Duplicated stimuli is inappropriate for behavioral as well as EEG paradigms, which generally aim to avoid identical word repetition (unless in a 'repetition' condition). Accordingly, these sets of word pairs were reduced to sets containing only unique words. The final set of words totaled 345 pairs. Four pairs were later identified as containing duplicates with the remaining set, and were removed, leaving 341 pairs. During data collection, five word pairs that should have been rejected during the taboo word screening were identified. These

[10] Due to a typo in the author's code to generate the LSA model, these LSA values are based on a 30 dimensional model rather than a 300 dimensional model. This typo was discovered after human subjects data collection but before data analysis, so all later LSA values used in the analyses are from the (correct) 300-dimensional model. This error is not a major concern because the purpose of using LSA during stimuli selection was to group stimuli into very general bins of similarities, so precise assessment is not crucial. Additionally, the two versions of the model correlate with a Pearson correlation of 0.628 and Spearman correlation of 0.716.

words were changed to non-taboo words for the remaining participants and the five involved pairs were rejected post-hoc. Final analyses were conducted on 336 word pairs.

4.1.2 For model predictions in Experiment 1b

The stimulus set described above was constrained in size due to the needs of human participants. If no humans are involved, or if pre-collected human data is used, then the stimulus set can be quite large. To expand upon some of the stimuli in the set described above, we tested the GOLD model and LSA on the complete sets of word pair stimuli from Plaut & Booth (2000) and Chiarello et al (1990). Plaut and Booth's 240 word pairs are categorized as related and unrelated, based on free association norms (Nelson et al., 1999). Chiarello et al.'s 144 word pairs are sorted into three categories according to relationship type: associated only, similar only, and word pairs that are both similar and associated. These categorizations were assigned based on several sets of norms, and the words were balanced on length, frequency, and imageability.

It is worth noting that some of the stimuli from Chiarello dated themselves; ostensibly related pairs such as *decoy-duck* were rated as unrelated by all participants in Experiment 1a, suggesting that this pair is no longer reliably associated in the modern lexicon. The same may be argued of some of the older commonly used sets, such as Rubenstein and Goodenough's set (1965) that includes terms with vulgar connotations in modern parlance. Accordingly, post-hoc sorting and plotting of ERP data that was collected in Experiment 2 was based on rating data as well as predefined word categories, as the rating data may better reflect the lexicon and language experience of the ERP participants.

4.2 Participants (1a)

Reaction times and judgment data were collected in two tasks: the first was a task of similarity judgment, and the second a task of association judgment. Participants were 34 undergraduate students (3 male) in the association task, and 31 undergraduate students (7 male) in the similarity task, recruited from the Psychology Department participant pool and compensated with course credit. All were native English speakers. None of the participants who contributed data to the word pair judgment tasks also contributed data to the ERP task.

4.3 Procedure (1a)

In each of the tasks, participants gave informed consent and then were seated at a standard desktop computer. Participants were first instructed on the nature of the relationship they were to judge, and then completed several example trials with the experimenter, discussing their judgments on each example trial. After the experimenter was satisfied that the instructions were understood, the participant then completed 341 trials, self-paced. Each trial consisted of a word pair presented with a Likert scale (1-7) with ends labeled as maximally or minimally related based on the specific relationship in the task.

4.4 Data analysis (1a)

Brief post-hoc interviews with participants indicated some difficulty regarding task instructions, ranging from forgetting the instructions partway through the task to inconsistency in following task-specific instructions. Data were cleaned by removing trials whose RTs were below 500ms (36 out of 11,594 trials in the association judgment task, and 12 out of 10,571 trials in the similarity judgment task).

4.5 Results

4.5.1 Ratings (1a)

Rating data on the similarity and association judgment tasks were treated as continuous data and were separately predicted using several regression algorithms: support vector regressors (SVR), random forests, and k-nearest-neighbors. GOLD output and LSA were separately used as input features to these algorithms. Performance measures are averaged across 10 iterations of training and testing on randomly selected subsets of the data (70/30 train/test). Performance was quantified via r-squared and root mean squared error (RMSE), which is not meaningful alone and is thus compared to a predictor that always predicts the training set mean. The default parameters from the Orange software suite were used for each algorithm: SVM regression (type=nu, cost=8.0, complexity bound=0.5, kernel type=RBF, tolerance=.001), random forests (maximum 20 trees, minimum 5 numbers of instances per leaf), and k-nearest-neighbors (5 neighbors, weighting by Euclidean distance, normalizing continuous attributes).

Table 1. Regressor performance on similarity and association ratings. Highest performance for each model is in a red font.

		Association		**Similarity**	
	Algorithm	*RMSE*	r^2	*RMSE*	r^2
	Mean	*2.0308*	*-0.0173*	*1.6779*	*-0.015*
smallGOLD	*SVM Regression*	1.3869	0.5255	1.2273	0.4571

	Random Forest	1.2625	0.6068	1.1081	0.5575
	kNN	1.4437	0.4859	1.3023	0.3887
GOLD	*SVM Regression*	1.3163	0.5726	1.2025	0.4789
	Random Forest	1.2498	0.6147	1.1595	0.5155
	kNN	1.3336	0.5613	1.2709	0.4179
LSA	*SVM Regression*	1.6461	0.3317	1.3752	0.3184
	Random Forest	1.7227	0.2679	1.4082	0.2853
	kNN	1.9561	0.0562	1.5906	0.0881

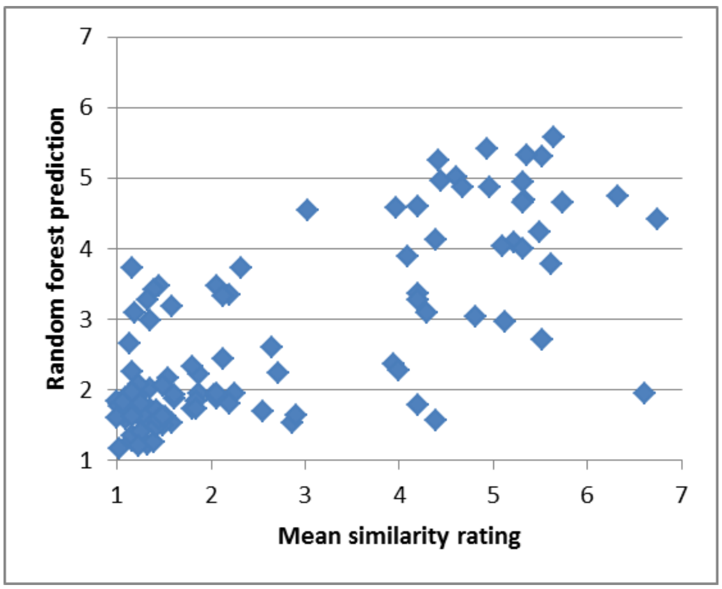

Figure 4. Similarity predictions from one train/test using a random forest trained on smallGOLD (r=0.75).

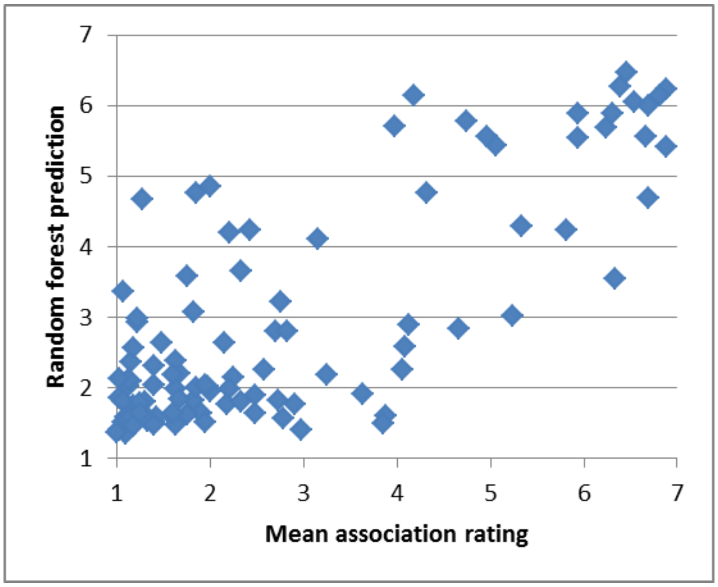

Figure 5. Association predictions from one train/test iteration using a random forest trained on smallGOLD (r=0.79).

GOLD and smallGOLD performed roughly equally, and quite well, at the task of predicting similarity and association ratings, with a maximum Pearson's r = 0.78. One set of train/test from each set of ratings was randomly selected for display in Figures 6 and 7. LSA did not perform as well at this task; to ensure a fair assessment, raw Pearson correlations were also calculated between LSA and association ratings (r = 0.5847, r^2 = 0.3418) and between LSA and similarity ratings (r = 0.5827, r^2 = 0.3395).

While GOLD performed well on the task of predicting continuous rating data, the high variability in human ratings suggests that these relationships may not all be 'true', in the sense that they are not agreed upon by multiple speakers. A subset of the word pairs judged in the above tasks were drawn from sets of words with predefined relationships, such as the words from Chiarello et al. (1990) which were categorized into words that were associated only, similar only, or both similar and associated. These predefinitions rest on datasets that may more reliably reflect the underlying word relationships, if at a coarser scale. Another set of words, from Plaut & Booth (1995), were categorized as related or unrelated, regardless of relationship type, which is at an even coarser scale. Accordingly, we next tested model performance on these full sets of words: first, the simpler classification task of related-unrelated pairs from Plaut & Booth (1995), and then the more complex task of distinguishing between the types of word relationships in the pairs from Chiarello et al. (1990).

4.6.1 Word pair categories (1b)

4.6.1.1 Distinguishing between related and unrelated words

Performance measures are averaged across 10 iterations of training and testing on randomly selected subsets of the data (70/30 train/test). Performance measures of accuracy, sensitivity (rate of true positives/'hits'), and specificity (rate of true negatives/'correct rejections') are presented, as well as confusion matrices. LSA was tested using several algorithms; best overall performance was achieved with neural networks (parameters: 1 hidden layer, 20 hidden layer neurons, regularization factor=1.0, maximum 300 iterations), so those data are presented here.

Table 2. Classifier performance on the Plaut and Booth (2000) word pairs.

	Accuracy	**Sensitivity**		**Specificity**	
		Related	*Unrelated*	*Related*	*Unrelated*
smallGOLD	0.9000	0.8914	0.9086	0.9086	0.8914
GOLD	0.9043	0.9000	0.9086	0.9086	0.9000
LSA	0.7443	0.6629	0.8257	0.8257	0.6629

Table 3. Classifier confusion matrices for the Plaut and Booth (2000) word pairs. Red percentages are the correct classifications.

		smallGOLD	
		Related	*Unrelated*
True	*Related*	89.1%	10.9%

class	*Unrelated*	9.1%	90.9%

		GOLD	
		Related	*Unrelated*
True	*Related*	90.0%	10.0%
class	*Unrelated*	9.1%	90.9%
		LSA	
		Related	*Unrelated*
True	*Related*	66.3%	31.1%
class	*Unrelated*	24.9%	82.6%

The two GOLD models demonstrated nearly identical, high performance (90% accuracy). Inspection of word pairs that were incorrectly classified reveal that the unrelated words misclassified as related were sometimes clear errors (*right-found*) but often perhaps related (e.g. *split-fight*, *yell-burst, treat-equal*,). GOLD failed to identify some clearly related word pairs (e.g. *horse-stall, great-super, take-bring, gives-share, slice-piece, glue-paste, right-wrong, live-death*). It appears that several of these pairs have more specific relationships than relatedness, including synonymy and antonymy. LSA performed well (74% accuracy); its most common error was to mis-classify related words as unrelated.

4.6.1.2 Distinguishing among relationship types

Having established that GOLD can distinguish related from unrelated word pairs, we turn to the task of distinguishing type of relatedness. As stated earlier, the distinction between association and semantic similarity is often a matter of degree as these factors are not orthogonal to one another. Thus, finding word pairs that are stronger in one dimension than the other or are stronger in both is a difficult task. Chiarello and colleagues (1990) have identified 144 such word pairs that are semantically related (*table-bed*) based upon category membership norms, associatively related (*mold-bread*) based upon free-association norms, and both semantically and associatively related (*aunt-uncle*). Following Lund, Burgess, and Atchley (1995, Experiment 3), we tested whether the metrics of the GOLD model could reliably classify these patterns of relationships and compared the results of the GOLD model to those of LSA.

Table 4. Classifier performance on the Chiarello et al. (1990) word pairs.

	Accuracy	Sensitivity			Specificity		
		Associated	*Both*	*Similar*	*Associated*	*Both*	*Similar*
smallGOLD	0.6023	0.6000	0.4857	0.7214	0.8250	0.7621	0.8172
GOLD	0.5791	0.6067	0.4429	0.6857	0.7250	0.7897	0.8517
LSA	0.3884	0.2667	0.5857	0.3214	0.7643	0.6862	0.6345

Table 5. Classifier confusion matrices for the Chiarello et al. (1990) word pairs. Red percentages are the correct classifications.

		smallGOLD		
		Associated	*Both*	*Similar*
True class	*Associated*	60.0%	24.7%	15.3%
	Both	30.0%	48.6%	21.4%
	Similar	5.0%	22.9%	72.1%

		GOLD		
		Associated	*Both*	*Similar*
True class	*Associated*	60.7%	24.0%	15.3%
	Both	41.4%	44.3%	14.3%
	Similar	13.6%	17.9%	68.6%

		LSA		
		Associated	*Both*	*Similar*
True class	*Associated*	26.7%	27.3%	46.0%
	Both	15.0%	58.6%	26.4%
	Similar	32.1%	35.7%	32.1%

Overall accuracy is best for the smallGOLD model. Inspecting the confusion matrices indicates that the GOLD models' most common error is to mis-classify word pairs that are both similar and associated as associated-only; the next most common mistake is the reverse, where associated-only word pairs are mis-classified as both similar and associated. LSA's most common error is to mis-classify the associated-only words as similar-only. It also assigns similar-only words equally often to the three categories.

4.6.1.3 Feature analysis

This initial exploratory testing of the GOLD model relied on the 'shotgun approach' of feature generation, in which all of the combinations of normalization and metric calculation were used as inputs to the neural network. In order to determine which features the algorithm is relying on to produce its classifications, and perhaps to suggest which types of information are important for judging these word relationships, we investigated feature relevance using one- and two-feature classifiers, as well as standard feature selection methods. For the one- and two-feature classifiers, a neural network learner classified the similar/associated/both word pair on 5 iterations of 70/30 train/test splits. In the first round of analysis, the neural network was given each of the 105 smallGOLD features individually; maximum accuracy of the 105 classifiers reached 50%. The full set of 105 features was sorted and the 50 highest-accuracy features were retained. In the second round of analysis, the neural network was given all combinations of two features from these 50 features, one pair of features at a time; maximum accuracy reached 63% accuracy, which is on par with the full set of features. Inspection of these feature pairs revealed that the majority of the top ranked pairs included two types of metrics: Method 5 from the similarity metrics (which considered only overlapping nodes, weighted by magnitude difference and normalized by size) and the PMI calculation of association. The top 30 performers were all pairs that included one association and one similarity measure.

Limiting the neural network to those two methods (30 features) yielded 63% accuracy. Limiting the neural network inputs to those two metrics (30 features) yielded 63% accuracy. Using additional feature selection (linear SVM weights) to reduce the number of features to 10 produced 65% accuracy; reducing the number of features to 5 boosted accuracy to 68%, which is well in excess of performance using the full set. However, these performance outcomes should be interpreted as exploratory only. The broad conclusion regarding features is that the combination of association (direct connections between the two words) and similarity (based on the overlapping and nonoverlapping neighbors of the two words) metrics is more powerful at predicting category than either alone. It may be possible to conclude that the similarity metric considering normalized overlap only and the PMI calculation of association are the most useful, but the similar/associated/both word pairs are not designed to span the language space and thus this finding may not generalize to other regions of the graph.

Chapter 5: Experiment 2 (neural data)

5.1 Participants

Participants were 20 graduate and undergraduate students recruited from the University of Maryland campus. Participants (7 male, 13 female; mean age = 25.15

and SD = 2.79) were all right-handed. One male participant's data were not considered in analyses, due to scores far below the sample mean on all of the reading and language assessments. All participants gave informed consent and were compensated for their participation with snacks.

5.2 Procedure

In the first hour of the study, participants completed the Peabody Picture Vocabulary Test (PPVT; Dunn & Dunn, 2007), both subtests of the Test of Word Reading Efficiency (TOWRE; Torgensen, Wagner, & Rashotte, 1999) the Nelson-Denny Vocabulary and Comprehension tests (Brown, Fishco, & Hanna, 1993), and a handedness questionnaire. All assessments were pencil-and-paper. The PPVT is a standardized measure of receptive vocabulary in which participants must identify pictures that represent the meanings of orally presented words. The TOWRE consists of two subtests: Sight Word Efficiency and Phonetic Decoding Efficiency. The Sight Word Efficiency subtest is a measure of word reading fluency in which participants must read a list of words in 45 seconds, emphasizing both speed and accuracy The Phonetic Decoding Efficiency subtest is a measure of phonemic decoding skill in which participants read a list of pronounceable nonwords (e.g. *pelnador*) in 45 seconds, again emphasizing both speed and accuracy. The Nelson-Denny comprises a multiple-choice vocabulary test and a comprehensions test in which participants read passages and answer questions based on those passages. These assessments were not analyzed in the following work, but were rather used to ensure that participants were high-skill readers. The mean performance of the 19 participants who contributed ERP data is presented in Appendix B.

Following these behavioral measures, participants were fitted with the EEG cap and electrodes, seated in front of a standard LCD monitor, and asked to place their right hand on the number pad of the keyboard. Responses were made using the ‘1’ and ‘2’ keys on the number pad, and the next trial advanced using the ‘enter’ key on the number pad as well, all with the right hand. Experimental trials proceeded as in Figure 8 below. Each trial began with a fixation cross in the center of the screen for 450-550ms, jittered. The first word of the pair appeared for 800ms, followed by a blank screen for 200ms; then the second word of the pair appeared for 800ms, followed by a blank screen for 1000ms, followed by a prompt to judge if the pair was related or unrelated. The prompt remained onscreen until the participant responded. Between trials, a neutral screen encouraged participants to blink as needed before pressing enter to begin the next trial. Participants were encouraged to rest if their EEG appeared to be showing higher alpha power, if they appeared drowsy, or at their own discretion. Each participant completed all 341 trials in roughly 30 minutes.

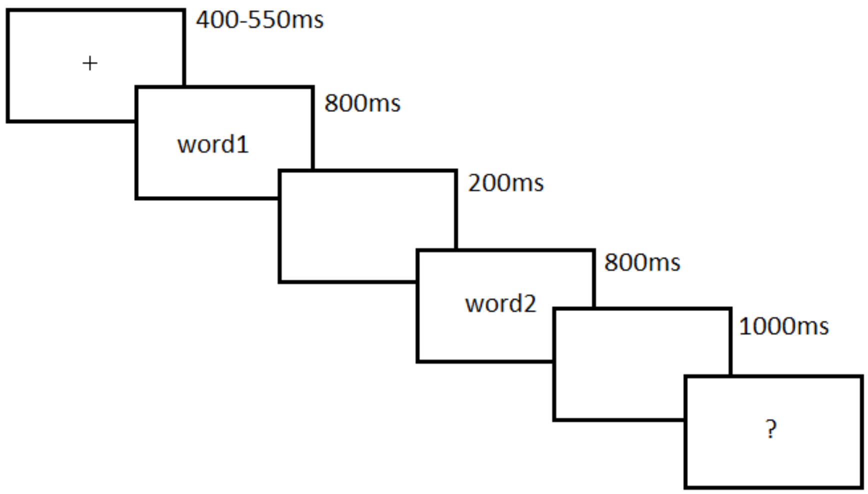

Figure 6. Trial template in the ERP task.

5.3 Data collection and analysis

5.2.1 ERP collection and preprocessing

EEG data were collected during the above task using the Biosemi system with a 64 channel electrode cap, referenced to linked mastoids. In two participants, one mastoid was irrecoverably noisy and/or separated from the scalp and thus their data were referenced to a single mastoid. In cases where a single scalp electrode failed (1 subject), it was interpolated. No more than one electrode was interpolated on any subject. No eye leads (EOG) were used; instead any trials contaminated by blink artifacts were rejected entirely. EEG was epoched (-200ms to 800ms), filtered (0.1Hz to 30Hz), and individual epochs rejected based on automated artifact identification (sliding window average). Trials were grand averaged by (a) word or response characteristics, discussed below with visualizations, and (b) by individual word pair, to be exported for per-stimulus ERP values.

5.2.2 Features for machine learning

A problem encountered in the course of ‘predicting neural activity’ is deciding what, exactly, should be predicted about neural activity. In the present study, the 64 channel electrode cap measured 512 timepoints per electrode per trial, which yielded ~30,000 data points per trial. It is reasonable to expect that only those timepoints and electrodes where the effect of word relationships is present will be predictable, so the tens of thousands of data points from other electrodes and time windows are not appropriate to consider. The n400 is typically measured as an average over the 300-500ms time window, and that the component is typically maximal over centro-parietal sites (Lau, Phillips, & Poeppel, 2008), so the present study restricted predictions to the average in the n400 window at the Pz and CPz sites.

5.3 Results

5.3.1 ERP visualizations and sanity checks

Grand average ERPs were visualized by averaging across trials sorted into various conditions in several ways: first, by individual subject responses (the ‘yes’ or ‘no’ judgments rendered while ERPs were collected); second, by the behavioral rating data in the relatedness and similarity tasks; and third, by category as defined in previous literature (the subset of words that appeared in the Chiarello et al. 1990 paper). As a sanity check, the first words of the word pairs in the yes-no judgment figure were plotted as well, to ensure no pre-existing differences that might reflect any number of errors.

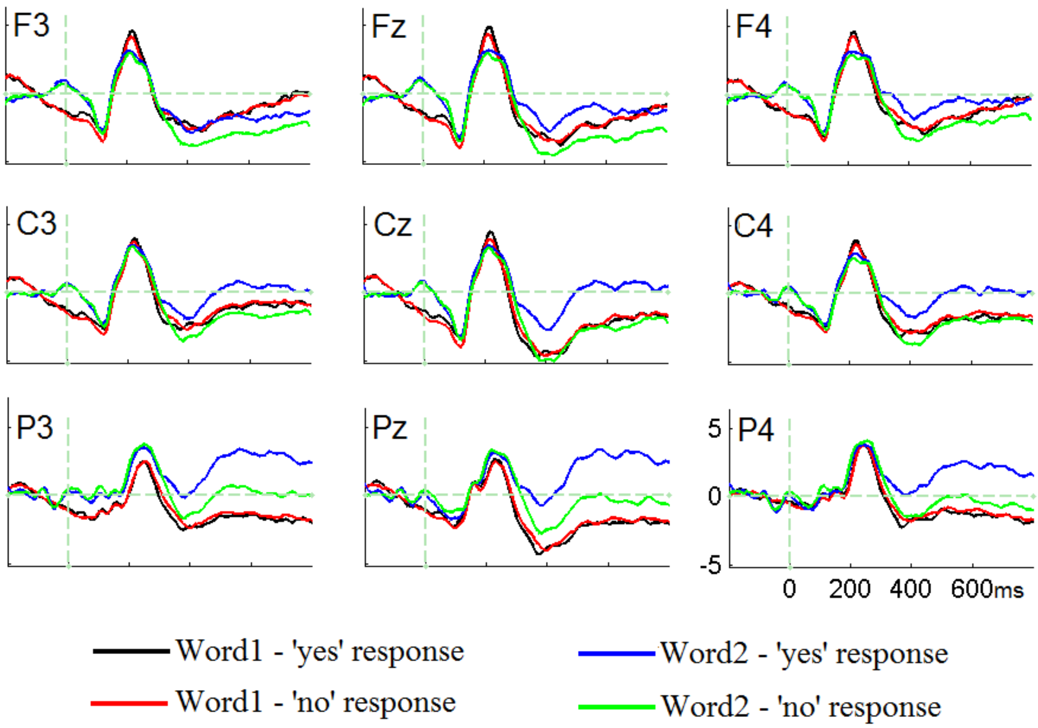

Figure 7. First and second words of the wordpairs, sorted by participant response.

Figure 9 above displays words that participants rated as related (‘yes’) and unrelated (‘no’). The first and second words of the word pair are displayed. Both word1s show a strong negativity in n400 window, which is to be expected, and are almost identical. Differences between the ‘yes’ and ‘no’ responses appear in the second word of the word pairs; related words produced an attenuated n400 compared to the first words of the pairs, and unrelated words produced either no difference or a smaller attenuation. This figure is assurance that the paradigm worked as intended in the broadest sense, and that the ERPs are thus far consistent with the literature.

The next set of figures will visualize the ERP data in several ways, and conduct statistical sets on certain contrasts. First, the ERPs sorted according to word pair rating will be presented and analyzed; then ERPs sorted according to category (the word pairs from Chiarello et al., 1990) will be presented and analyzed.

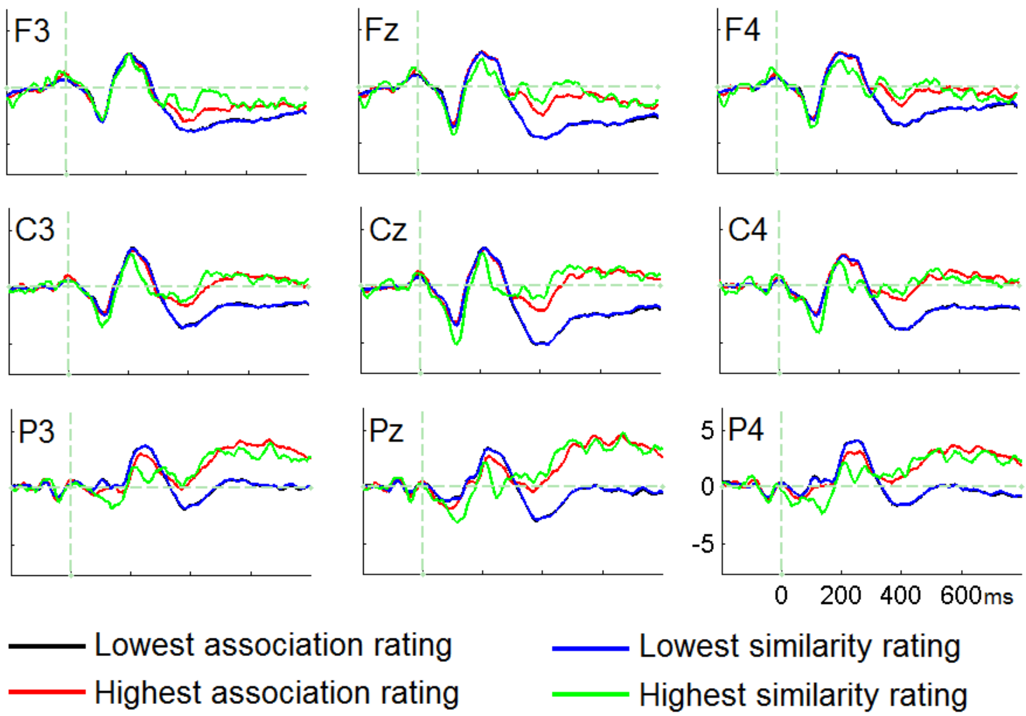

Figure 8. Second words of the word pairs, sorted into high and low similarity and association ratings.

Figure 10 above shows the second words of the word pairs, sorted into bins according to their ratings (by a different set of participants, in Experiment 1a). However, each trial contributes to two bins in these visualizations (each pair has both a similarity and an association rating), and many word pairs that were rated as minimally associated were also rated as minimally similar, so the two traces that look nearly identical *are* nearly identical, because they comprise a nearly identical set of

ERPs. In this figure it appears that words with the lowest ratings produced a large n400, and that highly rated similar and highly rated associated words each produced an attenuation of the n400 compared to their lower-rated counterparts. To examine this in more detail, Figure 11 and 12 present trials sorted by ratings binned into 6 bins, where each bin spans a single interval of the 7-point Likert scale (e.g. bin 1 holds word pairs rated from 1 to 2, bin 2 holds word pairs rated from 2 to 3, etc.).

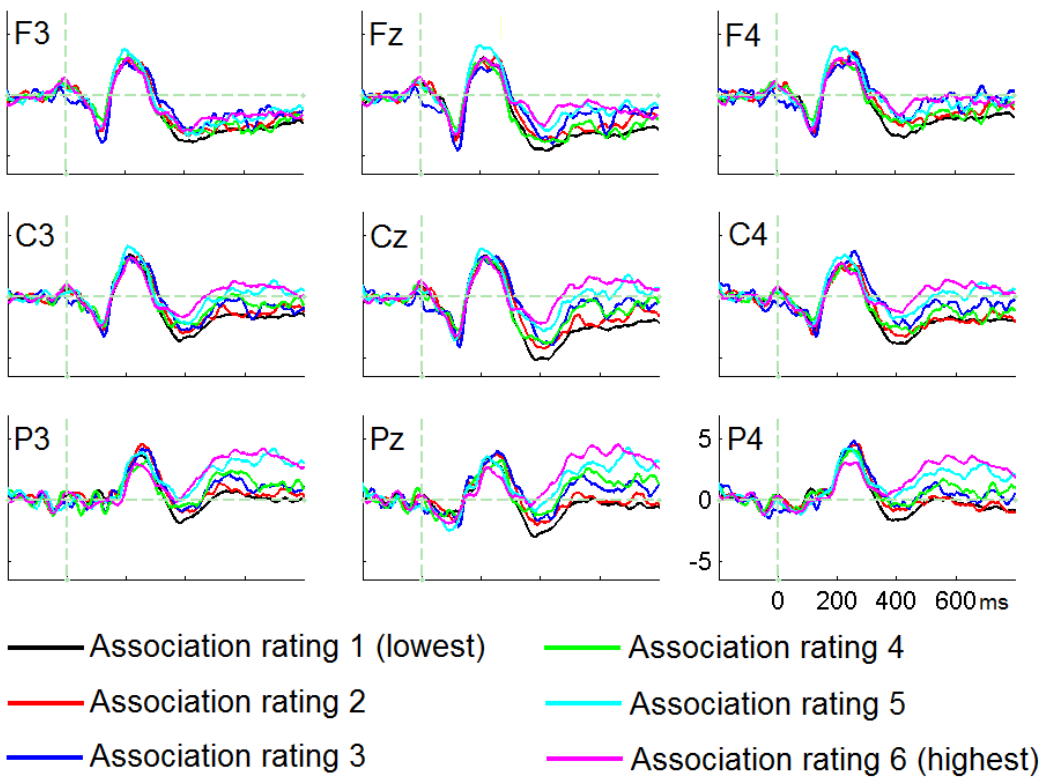

Figure 9. ERPs sorted by association ratings in six ordered bins.

Figure 11 shows the traces for the association ratings, divided into six bins. Across sites, but particularly clearly at Pz, the magnitude of voltage dip in the n400 window appears to be modulated by the degree of association.

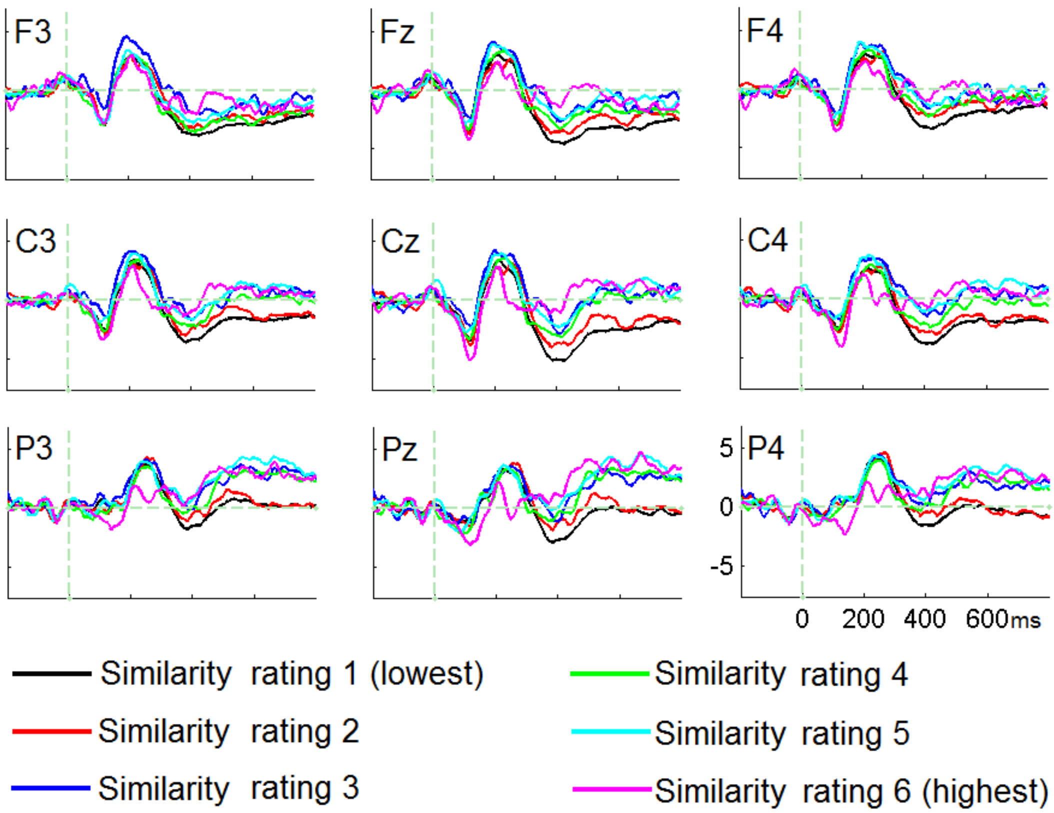

Figure 10. ERPs sorted by similarity ratings in six ordered bins.

Figure 12 is as Figure 11, but displays bins of similarity ratings rather than association ratings. The modulation of the n400 by degree of similarity is still apparent but less clear. This may reflect a genuine effect of similarity, or it may be the case that the range of similarity in the present stimulus set is smaller or differently distributed than the range of association. However, this and the previous figures plotted only mean waveforms and included no variability information and no statistical tests.

To determine if the ratings are reflected by real differences in the ERPs, statistical analyses were conducted on the highest versus the lowest bins of each of similarity and association, using t-maps or raster plots produced using the cluster-based permutation test from the Mass Univariate ERP Toolbox (Groppe, Urbach, &

Kutas, 2011). Cluster-based permutation tests capitalize on the broadly distributed effects of interest as well as the spatial density of the 64-channel electrode array. Additionally, although there are clear a priori predictions regarding the spatiotemporal distribution of effects for highly similar words, it is not known how these effects may change spatially or temporally with other types or degrees of relationships, and thus testing the entire timecourse and all electrodes using the cluster-based permutation test is appropriate (Groppe, Urbach, & Kutas, 2011). Raster plots were produced with the Mass Univariate ERP Toolbox. The raster plots display electrodes on the vertical axis (upper set is left hemisphere, middle set is midline, and lower set is right hemisphere; within each set, moving from top to bottom moves from anterior to posterior), and time on the horizontal axis. Filled electrode x timepoint boxes represent spatiotemporal locations with a significant difference (white boxes = condition 1 is more positive than condition 2, black boxes = condition 1 is more negative than condition 2).

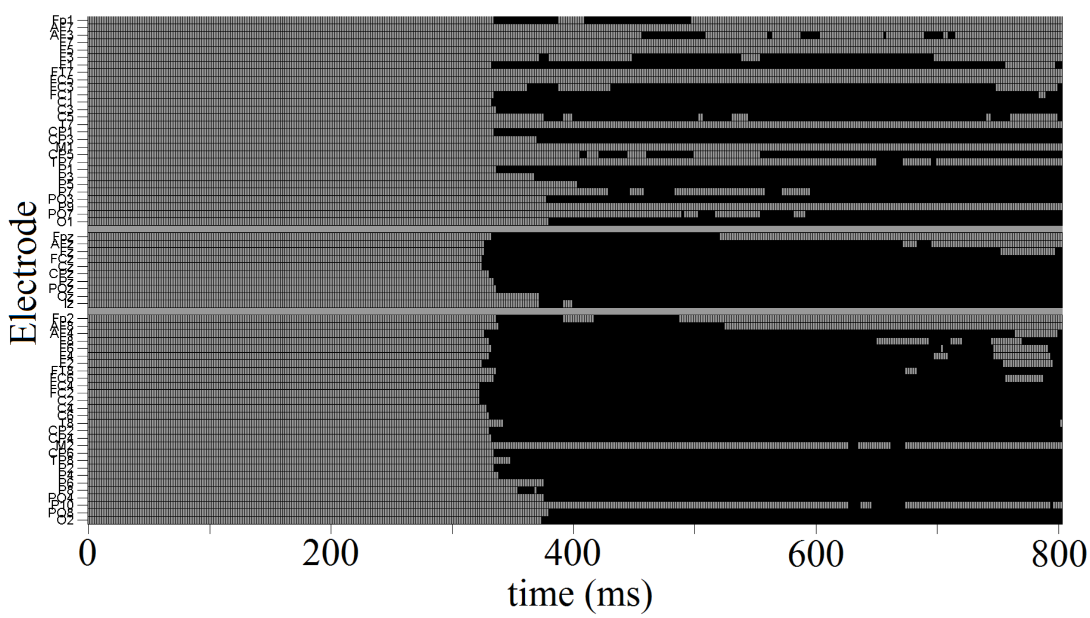

Figure 11. Main effect of association (lowest-highest).

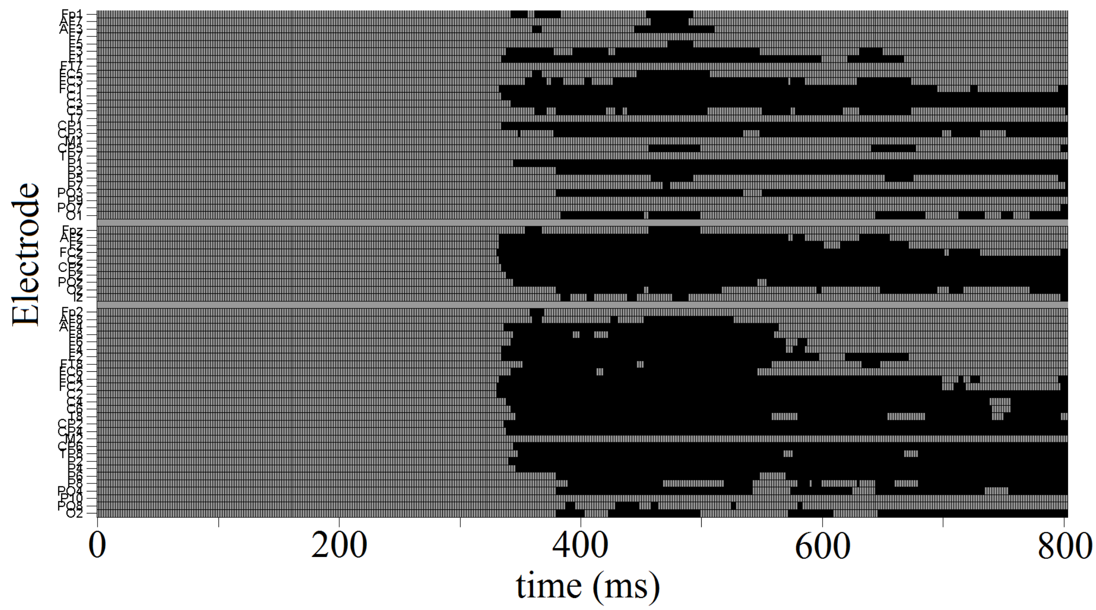

Figure 12. Main effect of similarity: lowest-highest

Figure 13 shows an n400 effect of association arising at around 300ms and extending through the rest of the epoch. Figure 14 shows and n400 effect of similarity, also arising at around 300ms and extending through the rest of the epoch. To determine if the spatiotemporal distributions of these two effects, are different, the interaction was tested as well (figure not shown). It was not significant at any timepoint: the two effects arise at the same time, taper off with the same general timescale, and are broadly distributed across electrodes. Some studies have found differences in spatial or temporal distribution of association and similarity effects (Koivisto & Revonsuo, 2001), but this finding was not replicated in the present ratings data. The next section examines ERPs to these word relationships sorted by predetermined category, rather than ratings.

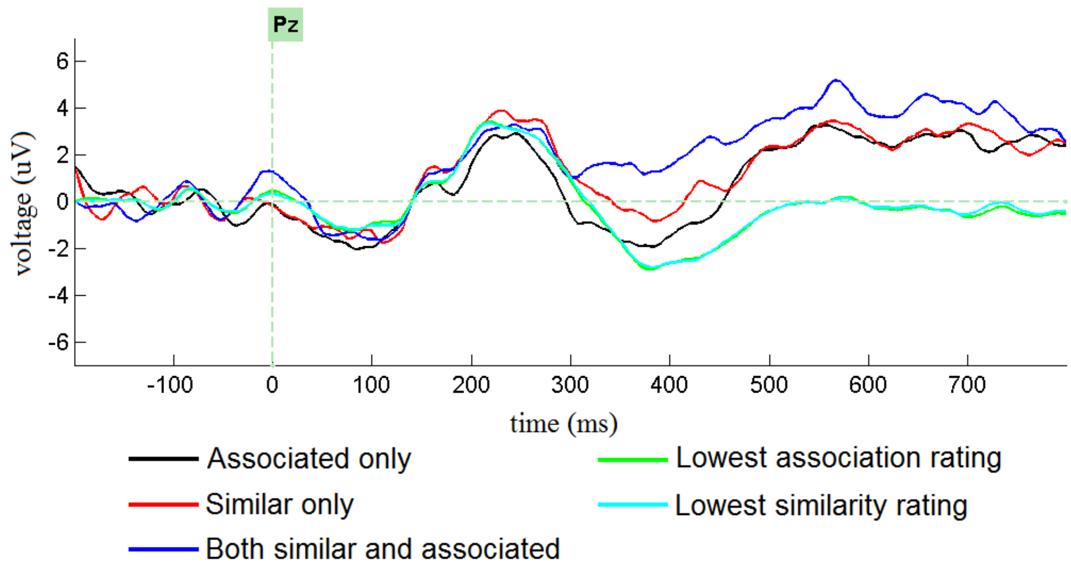

Figure 13. Chiarello et al. (1990) words vs. lowest rated words.

Figure 15 above displays the Chiarello et al. (1990) associated, similar, and similar-and-associated words compared to the words with the lowest ratings. All of the Chiarello et al. (1990) words produce some degree of attenuation of the n400 of the lowest rated words, but the degree of association appears to be graded. Words with both types of relationship produce the smallest n400, similar words produce a larger n400, and associated words produce an even larger n400.

To determine if these categories are reflected by real differences in the ERPs, statistical analyses were conducted on the three main effects of similarity, association, and both, as well as the interactions between these effects, using the cluster analysis described above. For present purposes, the word pairs rated lowest are referred to as ‘unrelated’ and are used as a baseline to which the categorically related words may be compared.

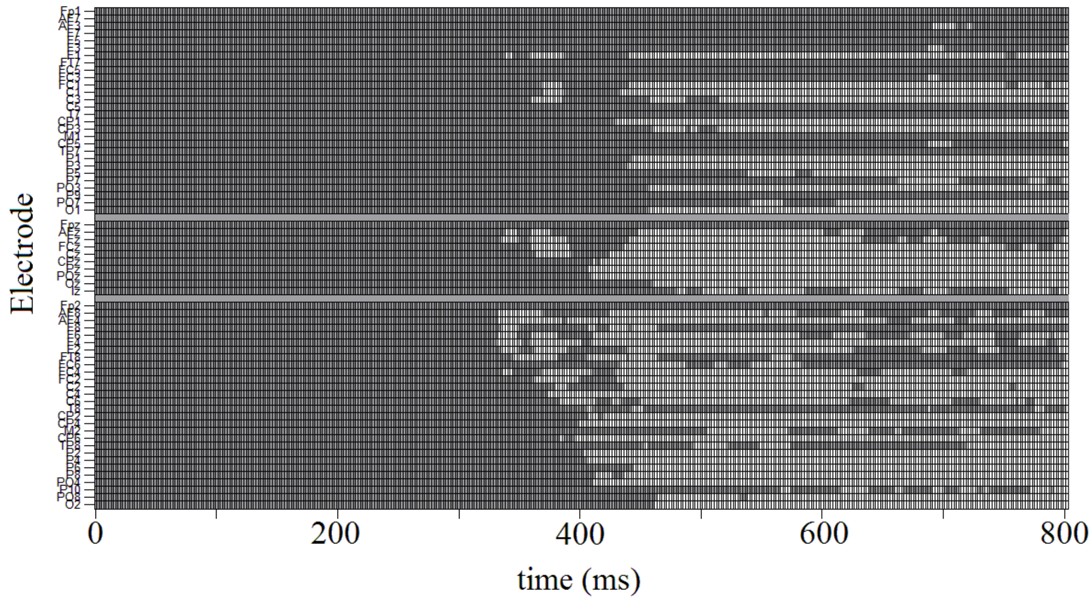

Figure 14. Main effect of association (associated-unrelated)

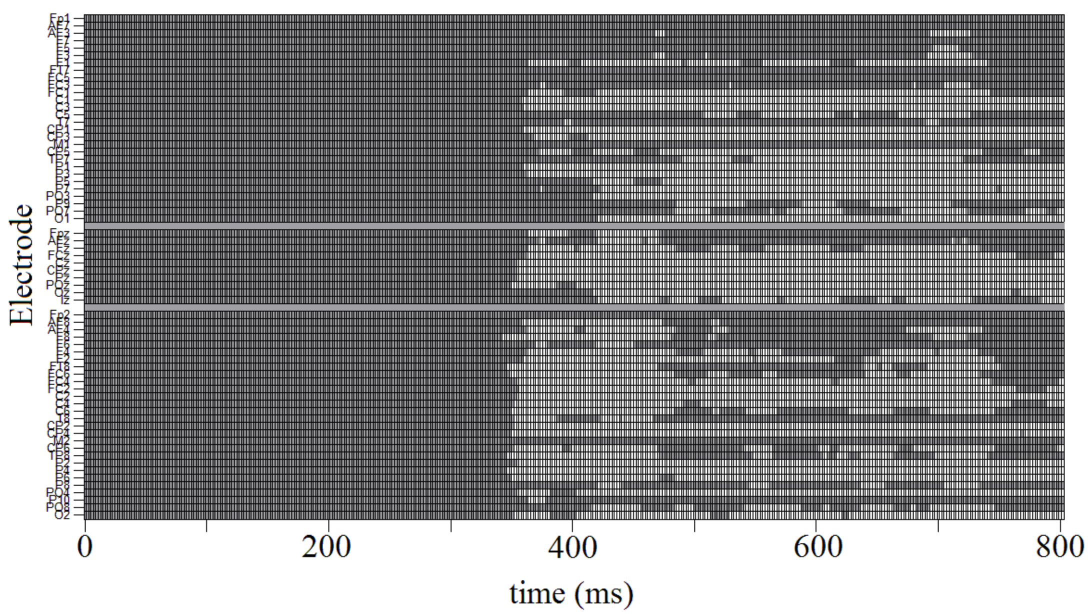

Figure 15. Main effect of similarity (similar-unrelated)

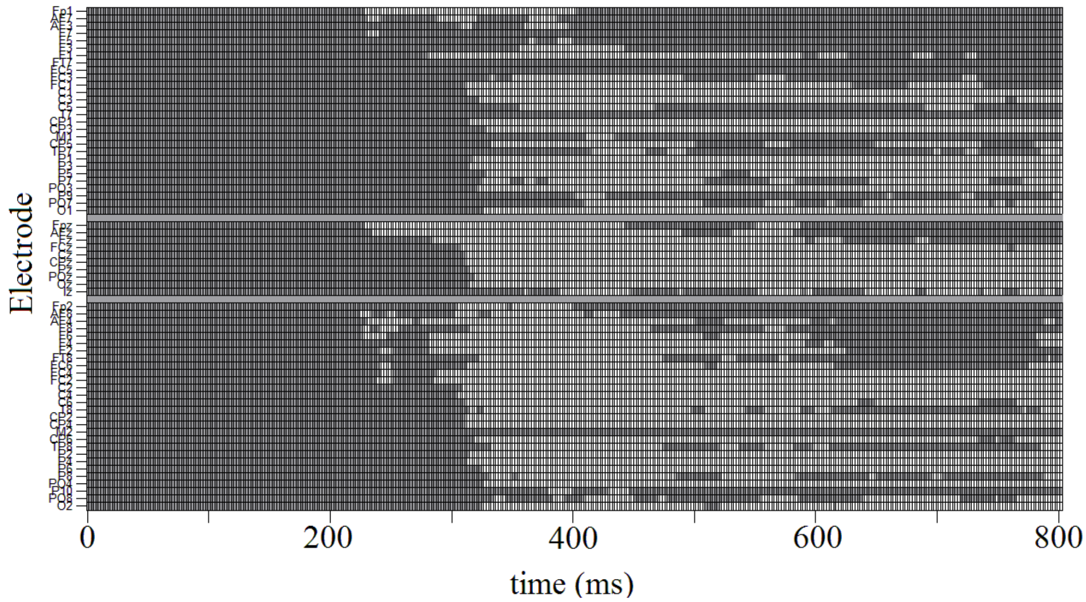

Figure 16. Main effect of similarity and association (both-unrelated)

Figures 16, 17, and 18 demonstrate main effects of the associated, similar, and both associated and similar relationships. In all three main effects, an n400 attenuation appears by roughly 300 or 350ms, such that the related words are more positive than the unrelated words, and lasts for the duration of the epoch. These rasters do show some variability, so the next section will present the interactions to test if the effects of each relationship type are different.

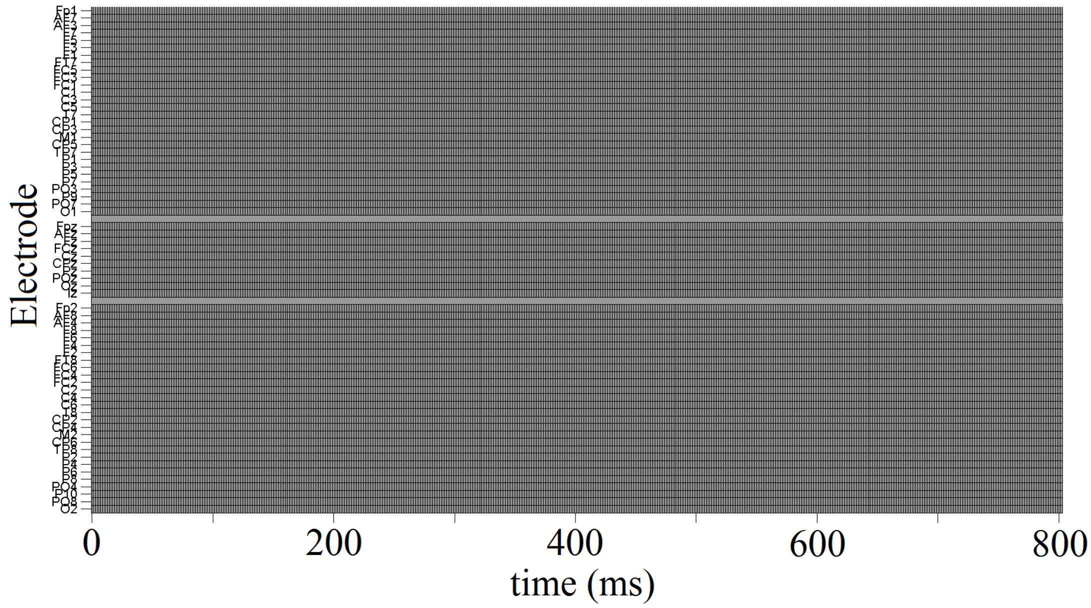

Figure 17. Interaction between association and similarity (associated-similar).

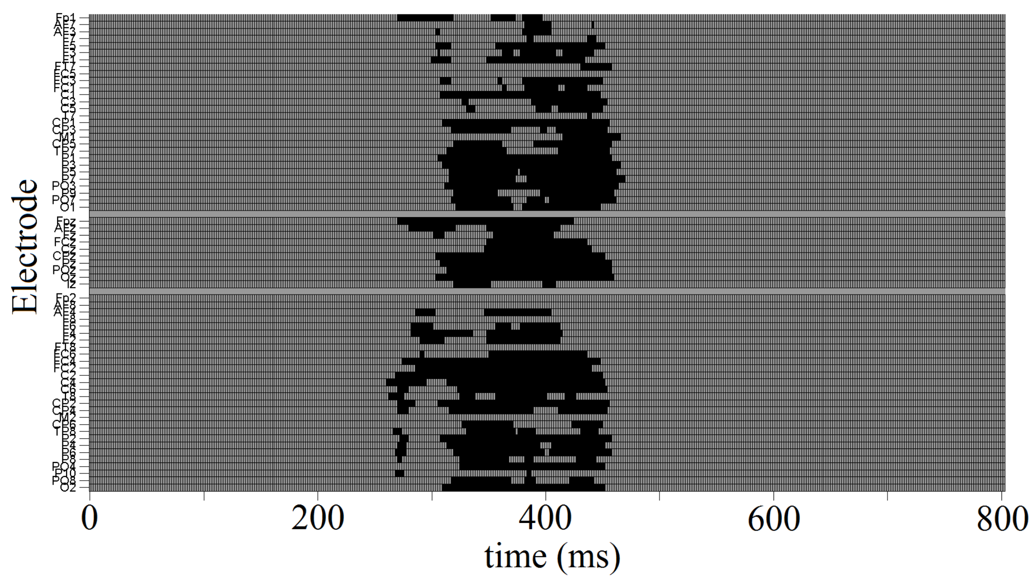

Figure 18. Interaction between association and both (associated-both)

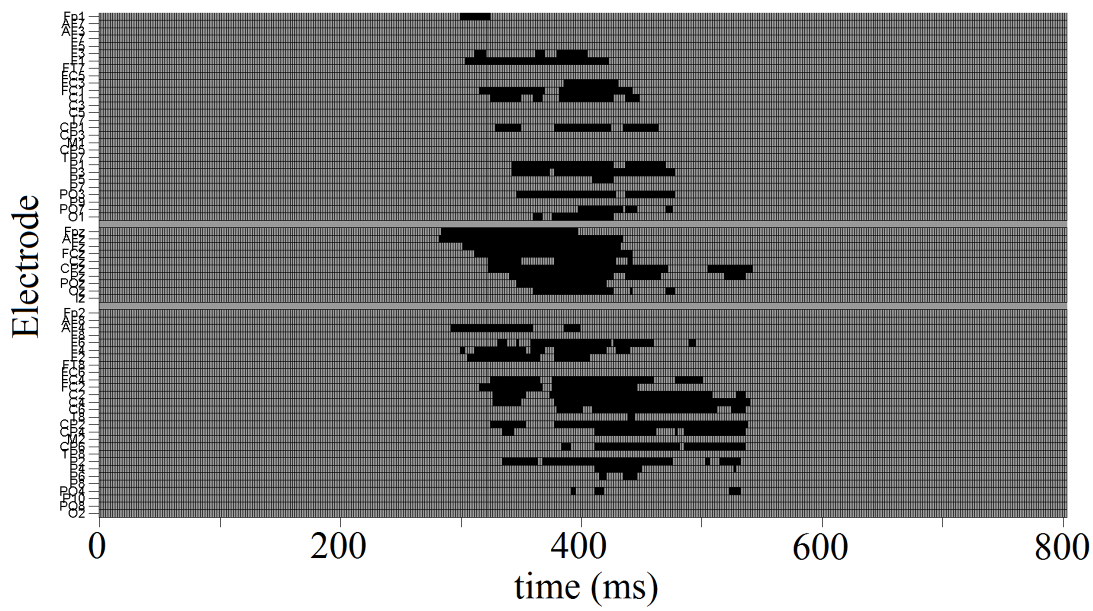

Figure 19. Interaction between similarity and both (similar-both)

Figures 19, 20, and 21 reveal that the interaction between the effect of similarity and the effect of association is not significant anywhere, but similarity and association each produce a smaller attenuation than both relationships together in the classic n400 window (300-500ms). These data support an account that the total relationship between two rods produces a particular n400 magnitude, rather than similarity or association contributing unique variance to the n400 magnitude.

However; of the entire set of 341 word pairs that neural data were collected , only a small subset were drawn from the Chiarello et al. (1990) pairs (30 associated only pairs, 23 similar only pairs, and 21 similar and associated pairs). The author has previously found significant n400 effects and interactions with a similar number of trials per condition on the same hardware, software, and workflow, and with similar participants (Jackson & Bolger, in preparation), but, in the present study, it is possible that certain effects are present but would only reach significance with a larger pool of trials per participant. However, the choice of analysis (cluster analysis using the Mass

Univariate Toolbox) gives a high probability of finding an effect if it is large, which n400 effects tend to be. In summary, it is possible that a difference between similarity and association would be apparent in ERP under different circumstances.

All of these visualizations demonstrate a clear n400, followed by a difference that lasts throughout the remainder of the epoch at a subset of the electrodes. This is not a common finding in the ERP literature, but it is a pattern that we have observed in language tasks recorded on the same equipment with a similar pool of subjects in the past. Whether this extended difference represents a genuine finding or an error of some sort in collection or processing is not clear. However, for the present, analyses will be confined to the n400 window, in which these ERPs display a canonical form.

In summary, initial examinations of the ERPs are generally consistent with previous literature. Similarity and association are both reflected in the n400, though perhaps not differentially. We next turn to predictions of these ERPs.

5.3.2 Model predictions of ERP voltage

Average voltages in the n400 time window at Pz, averaged across subjects, were treated as continuous data and were predicted using several regression algorithms: support vector regressors (SVR), random forests, and k-nearest-neighbors. GOLD output and LSA were separately used as input features to these algorithms. Additionally, similarity ratings and association ratings from Experiment 1a were used as predictors (each individually, and summed) to determine if that information is sufficient to predict neural activity. Performance was quantified via RMSE and r^2 as in Experiment 1a, using the same algorithm parameters.

Table 6. Regressor performance on voltage at Pz, 300-500ms.

		Pz300-500ms	
	Algorithm	*RMSE*	*R2*
	Mean	2.0414	-0.0038
smallGOLD	*SVM Regression*	1.9999	0.0366
	Random Forest	2.1100	-0.0724
	kNN	2.3154	-0.2914
LSA	*SVM Regression*	2.0499	-0.0122
	Random Forest	2.2054	-0.1716
	kNN	2.5260	-0.5370
Ratings	*SVM Regression*	2.0271	0.0102
	Random Forest	2.1136	-0.076
	kNN	2.4171	-0.4073

Performance on this task was best in all cases using SVM, but the maximum performance achieved was smallGOLD's r^2 of 0.0366, which is unimpressive. It is particularly strange that the ratings produce such poor performance as well. However, note that several of the r^2 values are negative; this may indicate that r^2 is an inappropriate measure, perhaps due to nonlinearity in the ERP data (Tremblay & Newman, 2013). Following Carlson et al. (2014), Spearman correlations were calculated for one randomly selected set of train/test for each prediction method. To ensure that the machine learning methods did not detract from the performance that a raw correlation would produce, those correlations were calculated as well.

Table 7. Correlations between metrics and ERP measures.

	Pearson	Spearman
SVM-smGOLD	0.237	0.246
SVM-LSA300	-0.103	-0.101
SVM-ratings	0.209	0.157
LSAval300	-0.112	-0.099
AssocRating	-0.079	-0.059
SimRating	-0.062	0.023

As this single iteration of train/test may be a fluke, the correlations between predicted ERP values and true ERP values for the test sets of 20 iterations of train/test were calculated for smallGOLD, SVM-LSA, and the raw LSA values. The correlations are reported in full in Appendix C. Correlations between the true ERP values and the raw LSA values were slightly higher than the SVM-LSA values, so raw LSA was taken as the best LSA performance. A t-test assuming unequal variances (Ruxton, 2006) was conducted on the Spearman correlations for smallGOLD and LSA; this test and found a significant difference, $t(30) = 7.02$, $p < .001$, such that smallGOLD correlations (M = 0.228, SD = 0.084) were significantly higher than LSA's (M = 0.076, SD = 0.048).

In comparison to the better behavioral data predictions in Experiment 1, this may also seem unimpressive. However, it is important to note standards from the literature. To refer to a recent example of predicting neuroimaging data, Carlson et al. (2014) calculate Spearman correlations between various computational models and brain activity in two different brain regions; the maximum Spearman correlation that any of the models achieved was $\rho = 0.154$ (shown in their Figure 2). Accordingly, the mean smallGOLD performance of $\rho = 0.228$ may be acceptable.

Chapter 6: Discussion

6.1 Model performance

The fundamental goal of this paper was to demonstrate that as a computational model using more psychologically plausible architecture, the GOLD model could viably account for the relations between words using a graph constructed from the single mechanism of co-occurrences between words in discourse context. As such, the GOLD model performed very well (90% accuracy) on the simpler task of classifying words as related or unrelated. It performed well, but not as well (60%+ accuracy) on the more difficult task of determining whether the Chiarello et al. (1990) word pairs were similar, related, or both similar and related; however, this performance is considered with respect to an LSA model that reached only 39% accuracy on this task. GOLD reached ~60%, ~50%, and ~70% on the three relationship categories considered individually, and when it erred, it tended to err on word pairs in the 'both' category, which may reflect model error or may reflect greater strength of one or the other type of relationship. It was also much less likely to classify a word pair with only relationship type (associated only or similar only) as the other relationship type; if it erred on these word pairs, it was much more likely to categorize them as 'both'.

GOLD was able to predict human ratings of similarity and association with high accuracy as well (Pearson's r ranging from 0.7 to 0.8), again outperforming LSA's r = 0.58. The task of predicting brain activity was much harder for both GOLD and LSA, and even the human judgments performed poorly, as measured by r^2. However, an analysis based on previous literature that predicted neural activity from language models indicated both that Spearman's correlation is more appropriate given nature of neural activity, and that GOLD's performance was actually quite good in the context of prior findings. One potential source of difficulty in predicting the ERP measure is that even fine-grained behavioral ratings of word pairs on the similarity and association axes were poor predictors. It may be the case that the influence of similarity and association combine in some nonlinear fashion to produce the n400 that is ultimately measured, or it may be the case that another variety of relationship entirely is also contributing variability to the ERP. Additional, direct testing of the n400 did not show waveform magnitude differences based on the type of relationship of the words that produced it; if anything, the n400 magnitude appeared to reflect total amount of relationship rather than any specific subtype.

The predictive power of the GOLD model, which was constructed from co-occurrence alone, indicates that the information used to judge relationships among words may be present in lexical co-occurrence alone, without considering additional language information such as word order. Furthermore, because GOLD was able to predict multiple, graded varieties of relationships between words (similarity and association), it is implied that information sufficient to represent both relationship types is present in lexical co-occurrence. This predictive success lends support to a single-mechanism model of word knowledge, and suggests that the method of calculating relationships, rather than representing relationships, may be what differs between relationship types. This is consistent with theories that word meaning is constructed or retrieved on an ad-hoc basis (Kwantes, 2005, see Neely, 1991 for review), as multiple mechanisms of querying may reasonably be involved in that ad-hoc construction. Preliminary analysis of the neural network classifier using the GOLD metrics indicates that the combination of association and similarity metrics are more powerful predictors than either type of metric alone, which lends additional support to this multiple querying mechanism account of word meaning. However, the data predicted in the present study were not reaction time data, as from priming studies, that may better distinguish between relationship types, as was done in Lund, Burgess, and Atchley (1995). As such, GOLD is agnostic as to which specific processes (such as automatic spreading of activation or post-lexical retrieval processes) its predictions are modeling or may be reflecting.

6.2 Word relationships

An alternative explanation for GOLD's misclassifications may not reflect an error in the model, but rather the fundamental difficulty of assigning words to different relationship types, which are non-orthogonal categories, as Chiarello and colleagues (1990) have done. In essence, the GOLD model, using a corpus of more natural language use and preserving that history in the connectivity patterns, may reveal that conceptually related words co-occur more frequently than assumed on the basis of free association norms.

It may also be the case that the very question of "how similar are these two words" is ill-posed to some degree. Consider *hot* and *cold:* these words are antonyms, but both are temperatures, and thus perhaps more similar than *hot* and *rutabaga*. *Earthquake* and *tornado* are wildly different concepts, but in a list of *earthquake, tornado,* and *democracy,* suddenly they are much more similar. In this vein, is it even meaningful to ask if two items are similar in isolation, or is a larger context necessary? If the larger context is important, what is the brain actually doing with these word pairs in isolation? Clearly some sort of similarity judgment is possible, as an n400 response can be achieved in the case of minimal context, and furthermore, that n400 can be modulated by some manner of relationship between the prime and target words.

6.3 Benefits of computational models

As was discussed in chapter 1, it has been argued that computational models are merely tools, from which nothing of substantive value can be learned. The GOLD model and its performance in the present study are intended as an argument to the contrary: as a model of language, rather than a tool, GOLD produced evidence that supports specific theoretical accounts of language acquisition, word meaning, and the reflection of language in neural activity.

However, it is undeniable that computational models provide a major advantage in their capacity as tools, namely that computational models aren't people and thus are free of human foibles[11]. The model doesn't participate in the study inebriated, doesn't grow fatigued or fall asleep, doesn't ignore task instructions, and its performance doesn't change over time, all of which are problems that plague human subjects research. The ultimate effects of these foibles on research data fall into the categories of *consistency* and *following task instructions* (much akin to the duality of *accuracy* and *precision)*. For an example of both, during an informal post-hoc interview in Experiment 1a, one participant described that he "drifted into" judging a different aspect of word meaning partway through the twenty minute rating

[11] Model *construction*, of course, may be fraught with foible, but that is beyond the scope of the present study.

task; he had rated association for the first ten minutes, and then similarity for the last ten minutes. He was not consistent across word pairs in the session and was not following task instructions during the second half of the task. Other subjects encountered difficulties in following instructions, particularly in the semantic similarity judgment tasks, in which certain participants initially judged all word pairs as minimally similar because any two words in a pair "[were not] the same words". Certain studies have quantified within-subject variability on tasks of language judgment (e.g. Barsalou, 1987), and consistency varies widely; to the author's knowledge, no formal study has been conducted of participant noncompliance in language tasks of this nature. However, it is common practice in behavioral research to include questions whose answers are trivially easy (e.g. "Please fill box A on the response form for this question"), in order to check if participants are actually engaging with the task or following task instructions. In contrast to these problems, computational models perform with both accuracy and precision consistently and in a trivially replicable manner.

6.4 Graphs as models of language

Graphs are a valuable tool in psycholinguistics research, both in service of analysis and of understanding. As a boon to analysis, graphs do not require discarding vast tracts of data in the process of dimensionality reduction, and so the model may maintain a higher degree of complexity that preserves additional information about relationships between words as well as overall statistical regularities that reflect the model's 'experience' with language (see Steyvers & Tenenbaum, 2005). Analysis of a graph model of language rests on the centuries-old field of graph theory for a solid mathematical foundation and a broad array of analytical algorithms, which allow for assessment of structural as well as functional properties. These algorithms may be useful methods of modeling larger contexts in psychologically meaningful manners, through existing methods of modeling network propagation, etc. In terms of aiding understanding, graphs may allow for more intuitive interpretation of calculations and results than methods that require complex transformations of the data (e.g. SVD, Landauer, or circular convolution, Jones & Mewhort 2007).

However, these benefits, particularly the retained information, are accompanied by a major drawback: computational complexity. Analyzing graphs, particularly very large graphs as one might encounter in a language model, is computationally expensive. The patterns that may prove most interesting are also very complex; for example, identifying subgraph isomorphisms, one potential method of discovering useful patterns for word sense disambiguation or identifying word relationships, is in $O(|V_{graph}|^{|Vsubgraph|})$. Even performed in parallel, these operations quickly become intractable on standard hardware. Other types of graph theory algorithms may be valuable for identifying language features or word attributes, such as social network analysis for identifies ‘bridge nodes’ that may be homographs, or clique analysis that may be able to cluster register, or connotative/emotional content (Osgood, 1957), or feature similarities (McRae, De Sa, & Seidenberg, 1999; Plaut, 1995). These analyses are much more complex than something like LSA, and take exponentially more time to execute. The solutions to this complexity problem vary: recruiting massively parallel cloud computing resources, using only well-optimized algorithms and data representations (Sun, Wang, Wang, Shao, & Li, 2012), reducing the graph size, or just choosing analyses that can avoid the brute force approach.

One issue in graphs of word co-occurrence is that their high degree of interconnection makes many standard graph algorithms less useful, such as spanning trees and various measures of separation (e.g. Dijkstra, 1959). These algorithms are of course applicable, but may vary in their informativeness because the high degree of interconnectivity in a word-word graph means that words are typically very few steps away from any other word. In a graph like this, the weights of connections are more important than the presence of connections, so analyses must focus on algorithms that take weight into account, algorithms that consider larger patterns of weighted connectivity, or methods of graph pruning such that the presence of connections becomes informative – perhaps by pruning low weight connections, or limiting words to some arbitrary number of connections.

It may also be valuable to maintain more information during the graph construction process. In the present large GOLD model, each connection is weighted with weight=1, regardless of actual distance between words. It may be useful instead to record connection counts at several distances – e.g. *grumpy* and *cat* co-occur immediately adjacent n_0 times, separated by one word n_1 times, separated by two words n_2 times, etc. Maintaining word order information (perhaps through directional connections) may be a better predictor of human behavior as well, because, for example, *bread-butter* has a higher free association probability than *butter-bread*, etc.

Lastly, as with all models of language, vagaries of the corpus can influence model performance. The corpus from which the GOLD model in the present study was constructed may display a greater influence of conversational speech than, say, textbook-based corpora, as well as unorthodox grammatical structures and word usage. It also has a rather larger vocabulary of obscenities than a corpus constructed from the New York Times might, and spans different topics than standard language corpora (e.g. TASA; see Landauer et al., 1998). It was the aim of this corpus that it span a large range of unadulterated modern language use to again provide more ecological validity with respect to the behavioral data to which the GOLD model may be applied.

6.5 Individual differences

Individual variability in language experience (explored in the author's prior projects; (Bolger & Jackson, under review; Jackson & Bolger, in preparation) leads to dramatic differences in word knowledge and thus the neural response to words in context. In the case of paired priming paradigms, the context is minimal: one preceding word. Clearly, this minimal context is sufficient to bias the neural response, as the n400 effect may be reliably elicited in these paradigms. However, due to its brevity and low information density, this context may be less effective at preventing unrelated or idiosyncratic semantic activation than a sentence or larger preceding context might. For example: the pair *grumpy-cat* would elicit a small n400 from the author, who has encountered the feline referred to as Grumpy Cat[12] in digital form on many occasions, but a large n400 from someone who is unfamiliar with this animal. However, if the context were larger and contained more information and thus more constraint, such as "*the mouse toy was chewed up by the huge, orange, grumpy cat*",

[12] See www.grumpycats.com for details.

it may be the case that these two individuals' n400 responses to *cat* would be closer in magnitude.

The rating tasks in Experiment 1a provided a clear example of individual differences influencing word knowledge. The author presented *question-query* as an example of words that might be rated as highly similar; however, easily half of the participants rated this pair very low in similarity, because they had never encountered (or could not recall a meaning of) the word *query*. Incidentally, this is why participants with extensive vocabularies and high reading skill were selected to contribute the ERP data; the model should be predicting English in as complete or objectively accurate a form as possible, rather than being limited to modeling the smaller subset of language that is known to lower skill readers.

6.6 Future research

6.6.1 Language

The present study supports a single-mechanism account of the acquisition of these word relationships, but does not rule out an account in which acquisition is via a single mechanism, but later calculation or determination of the relationships (at time of judgment) occurs via multiple mechanisms. This question may be approached by examining the predictive elements of the model: are the features required for predicting association different than the features required for predicting similarity, and do these features reflect theoretical conceptions of association and similarity? Can the model predict other types of quantifications of word relationships, such as reaction time data, finer-grained ratings of word relationships, or neural activity in response to sets of words? Do sets of words constrain meaning and/or concept activation better than individual word primes?

6.6.2 GOLD

The present study explored whether the GOLD model could distinguish among similarity and associativity in word relationships. Future work should investigate whether GOLD can differentiate words along other axes and relationship types, such as antonyms/synonyms, multiple word senses, register, affective content, and so on. In support of these investigations should be the extraction of more complex measures from the graph, particularly those examining larger connectivity patterns. The present study was exploratory, and so was limited to an undirected, smaller graph and simpler, local algorithms. However, the full power of a graph model may lie in it its higher-order, more complex patterned relationships, so these should be evaluated.

Preliminary exploration of the ML algorithms used to predict activity and behavioral from GOLD does not make it obvious what is driving their obtained accuracy. It is not clear either way if either of the theoretically association-based (direct links between words) or the theoretically similarity-based (overlap and non-overlap between words' neighbors) metrics are more informative, or if the metrics are equally informative and the manner of weight normalization is more important. However, it is clear that combination of several features is more predictive than each feature alone. Further investigating what this may imply for human language processing will require a tightly controlled stimulus set that spans many axes of the language space.

A crucial element of future work will be the identification of optimal methods of prediction from the model. The present study used many features and machine learners to learn patterns that may be predictive; other studies have used such methods as scaling by arbitrary units (Lund & Burgess, 1996), and assessing predictive ability based on Spearman correlations (such as on dissimilarity matrices entries in Carlson, Simmons, Kriegeskorte, & Slevc, 2014, and on other types of data as in Collins-Thompson & Callan, 2007 and Gabrilovich & Markovitch, 2007, to name two of countless studies). It may be also the case that larger contexts, such as those already used in judgments of document similarity, are necessary for more meaningful judgments of similarity. Future research with the GOLD model should address the development of metrics from GOLD that can be expanded to arbitrary-length inputs, which may enable greater predictive power as well as more accurate modeling of psychological reality.

6.6.3 Individual differences

It is undeniable that individual differences contribute to neural responses to language. Future work may examine these individual differences by comparing neural activity in high-skill to readers to that in low-skill readers, particularly if the stimuli also vary along several dimensions of difficulty. The word stimuli used in the present study were fairly high frequency, but it’s not clear if higher-order interactions with words that are involved through spreading activation or other processes, or other additional information derived from greater experience with language, may have an effect on the measured waveforms.

6.6.3 ERP

One of the major goals of the present study was to predict brain responses in a language task. The present study used a very simplistic approach to quantifying these brain responses: average voltage in a specific time window of ERP at a single electrode. Unfortunately, this approach discards a tremendous amount of data that may be very relevant in terms of differentiating word characteristics or cognitive processes (e.g. Halgren et al., 2002; Sereno, Brewer, & O'Donnell, 2003; Thornhill & Van Petten, 2012). A different method of encoding the total spatiotemporal pattern of the brain response may be valuable to capitalize on the additional information present in such patterns.

Future work may also examine prediction in the other direction: predicting characteristics of words from ERPs. Using ERPs as predictors may better enable use of the entire spatiotemporal pattern of voltage, rather than collapsing such a complex pattern into a single value as in the present study. Koivisto and Revonsuo (2001) found that dividing the n400 window into early (250-375) and late (375-500) allowed for the discovery of different spatial and temporal patterns of effects for lexically associated as opposed to semantically similar word pairs; future work may follow this paper and attempt to predict differential activation in different time windows and electrode locations.

6.6.5 Extensions

In the interest of maintaining a sensible scope of the present project, these applications were not explored. However, these applications have clear relevance to the reading and language literature, the cognitive literature, and other work in the

Bolger lab. This chapter will identify and briefly discuss several potential applications that GOLD, ERP data, or behavioral data might address.

6.6.5.1 Context variability

The context variability hypothesis (Bolger et al., 2008) may be tested by replicating the contextual word learning paradigm (Jackson & Bolger, in prep) using GOLD as the ‘participant’. The model could be ‘taught’ novel words in the same way that human participants were taught: exposure to the novel words embedded in sentence contexts. Model performance on this task may be compared to the human data from Jackson & Bolger (in prep), which include multiple choice sentence completion, congruent/incongruent sentence judgments (including ERPs to this task), and participant-produced definitions.

6.6.5.2 Semantic distance in fMRI

Previous research in fMRI has found relationships between semantic distance of language input and activity in left IFG, bilateral MFG, and anterior temporal regions in a lexical decision priming task (Tivarus, Ibinson, Hillier, Schmalbrock, & Beversdorf, 2006), and in left frontopolar cortex in an analogy judgment task (Green, Kraemer, Fugelsang, Gray, & Dunbar, 2010). GOLD could attempt to predict activation from these studies.

6.6.5.3 Word sense disambiguation

Words can be ambiguous in different ways: polysemy refers to multiple related meanings (a *boot* on a foot and *to give something the boot*), while homonymy refers to multiple unrelated meanings (the *boot* on a foot and the *boot* of a car). Previous research has used various approaches, including clustering (Levin, Sharifi,

& Ball, 2006; Lin & Pantel, 2002; Widdows & Dorow, 2002), an information-based approach (Durda, Caron, & Buchanan, 2010), a second-order cluster approach (Schutze, 1998), Wikipedia-based methods (Gabrilovich & Markovitch, 2007; Li et al., 2011) that uses additional information in a query (e.g. river *bank* vs *bank* loan), and hybrid methods that use both distributional data and human-annotated knowledgebases (Jiang & Conrath, 1997; Marton, Mohammad, & Resnik, 2009).

GOLD may be able to disambiguate word senses based on the patterns of connectivity of the different senses. Bridge analyses, in certain social network analyses (Butts, 2008) and epidemiological modeling (Luke & Harris, 2007) aims to identify nodes that participate in otherwise disparate sub-networks of nodes (nodes that act as 'bridges' between groups). It may be the case that homonymous words are bridge nodes. For example, the word *ball* should be heavily interconnected with a group of nodes including *bat, throw, pitch, baseball, football,* which should all be heavily interconnected; *ball* should also be interconnected with a group that includes *gown, dance, gala,* and *invitation*, all of which should be heavily interconnected, none of which should be particularly heavily connected to the sport-related group.

This type of analysis may also be helpful in identifying where information was lost in the parsing process; for example, all input is forced to lowercase before being weighted, and accordingly the difference between *US* and *us* is not detected in the first-order structure of the graph. If bridge analysis identifies 'us' as participating in two largely disparate clusters, one centering around groups and the other centering around foreign policy and military exercises in the Middle East, then GOLD may be able to distinguish between these two words.

6.6.5.4 Synonymy

Distributional models generally perform well on tests of synonymy (Turney, 2001) and some methods have improved performance by specifically training on a thesaurus-based corpus (Jarmasz, 2003). Measures that preserve more dimensions are better at judging subtle differences between synonyms ("near-synonyms"), because less distinguishing information is discarded (Wang & Hirst, 2010). GOLD would not discard any data, and thus would be expected to perform well on a near-synonym judgment task (Inkpen, 2007; Turney, 2001), and may also be compared to human similarity judgments as in Budanitsky & Hirst (2005).

Theoretically, words with similar meanings should be connected in similar ways to other nodes. Standard cluster analysis (Hartuv & Shamir, 1999; Schaeffer, 2007) may be able to identify groups of words with similar meanings. The 'central' node – which measure of centrality would be appropriate here is an open question, but perhaps word frequency would be effective – would be the 'label' of that group. This could simplify further computations (by reducing many nodes to a single 'supernode'), or be useful for generative queries ('generate synonyms of *tired*').

6.6.5.6 Other

The model may be applicable towards a variety of other standard tasks, including authorship attribution, Cloze tasks, assessing metaphors, judging definitions, and so on. The model is further flexible in its parameters: by propagating activation through the network and manipulating parameters like falloff time and propagation rate, it may mimic parameters of human memory like WM span and speed of processing. Further work may address even more pie-in-the-sky hypotheses:

can the model suggest meaning for slang? Can it make rudimentary jokes, perhaps by completing an input sequence with a low-probability word?

6.7 Conclusion

The present study constructed GOLD, a graph model of language, from lexical co-occurrence, and used novel, theoretically-informed similarity metrics from GOLD to predict relationships among words, types of relationships among words, and neural activity elicited from reading words with particular relationships. The GOLD model is capable of distinguishing among types of relationships between words, predicting graded relationships between words, and predicting brain activity in response to words with varying relationships, using metrics constructed from theoretically-informed conceptualizations of association and similarity. These novel algorithms are theoretically informed in a straightforward manner: they consider how connections to associates that are common to both words and associates that are unique to each word differentially contribute to meaning. This type of calculation is more transparent in its reflection of the co-occurrence patterns of language that were used to construct the model than algorithms involving more complex transformations, and, because it doesn't rely on spatial relationships of word representations in a particular language space (e.g. cosine between two word vectors), may be better able to account for psycholinguistic properties that would not be reflected in orthogonal relationships in a vector space model.

Appendix A. GOLD metrics

Five methods were used to calculate similarity, all considering overlapping nodes and nonoverlapping nodes separately. It is theorized that a similar pattern of connectivity to overlapping nodes will arise when the word pair is more similar, but if their connections to nonoverlapping nodes are much greater, than the similarity in overlap may not contribute as much to the overall judgment of the word pairs. Accordingly, the following metrics involve various ways of summing weights to the overlapping nodes and summing weights to the nonoverlapping nodes, and comparing the two sums.

Method 1: Overlap and nonoverlap sets. The weights to each set are summed as follows, where |Vo| is the number of nodes in the overlap set, |Vn| is the number of nodes in the nonoverlap set, and $w_1 n_i$ is the weight between word 1 and node *i* :

$$Weights\ to\ overlap = \sum_{i=1}^{|Vo|} (w_1 n_i + w_2 n_i)$$

$$Weights\ to\ nonoverlap = \sum_{i=1}^{|Vn|} w_1 n_i + \sum_{i=1}^{|Vn|} w_2 n_i$$

However, any additive or subtractive combination of these values could be arbitrarily high. It would be ideal if the metric would map to a finite range for easy comparisons (like LSA's output ranges from -1 to 1). One approach is to compare the proportion of the total weights that is accounted for by weights to the overlap and the nonoverlap sets. The difference between these proportions will map from -1 (in the case where 100% of weights are connected to nonoverlap nodes) to 1 (in the case where 100% of weights are connected to overlap nodes).

$$Total\ weights = weights\ to\ overlap + weights\ to\ nonoverlap$$

$$Proportion\ to\ overlap = \frac{Weights\ to\ overlap}{Total\ weights}$$

$$Proportion\ to\ nonoverlap = \frac{Weights\ to\ nonoverlap}{Total\ weights}$$

$$Similarity = Proportion\ to\ overlap - Proportion\ to\ nonoverlap$$

Method 2: Overlap and nonoverlap sets, normalized by size. Method 2 is calculated as Method 1, except that $Weights\ to\ overlap$ and $Weights\ to\ nonoverlap$ are normalized by their relative sizes, as below:

$$Weights\ to\ overlap = \frac{\sum_{i=1}^{|Vo|}(w_1 n_i + w_2 n_i)}{|Vo|}$$

$$Weights\ to\ nonoverlap = \frac{\sum_{i=1}^{|Vn|} w_1 n_i + \sum_{i=1}^{|Vn|} w_2 n_i}{|Vn|}$$

The final similarity metric is calculated as in Method 1, as the difference of proportions to the overlap and nonoverlap sets.

Method 3: Overlap and nonoverlap sets, overlap set scaled by magnitude difference. For the remaining methods, the sum of weights to overlap transformed according to the following equation:

$$Weights\ to\ overlap = \sum_{i=0}^{|Vo|}\left(\frac{w_1 n_i + w_2 n_i}{\left(\frac{\max(w_1 n_i, w_2 n_i)}{\min(w_1 n_i, w_2 n_i)}\right)}\right)$$

This has the effect of scaling the two weights by how close they are in magnitude, such that weights that have a smaller magnitude difference will contribute more of their weight to the final total. In the example in Figure 3, *grumpy-face* has a weight of 9 while *cat-face* has a weight of 52; their combined transformed weight would be 10.56 (18% of the original combined weights). In contrast, *grumpy-depressed* has a weight of 2 while *cat-depressed* has a weight of 3; their combined transformed weight would be 3.33 (66% of the original combined weights).

In Method 3, weights to the overlap nodes are calculated as above, and the final similarity metric is calculated as in Method 1 (no additional normalization).

Method 4: Overlap and nonoverlap sets, overlap set scaled by magnitude difference, both sets normalized by size. In Method 4, weights to the overlap nodes are calculated as above and then normalized by size as in Method 2. The final similarity metric is calculated as in Method 1.

Method 5: Overlap set only, scaled by magnitude difference, normalized by size. In Method 5, only the overlap set is considered, and its weights are calculated as in Method 3 and normalized as in Method 2, as follows:

$$Weights\ to\ overlap = \frac{\sum_{i=0}^{|Vo|}\left(\frac{w_1 n_i + w_2 n_i}{\left(\frac{\max(w_1 n_i, w_2 n_i)}{\min(w_1 n_i, w_2 n_i)}\right)}\right)}{|Vo|}$$

Because the nonoverlap set is ignored, no proportions are calculated. This metric does not map from -1 to 1.

Table 8. Weight normalization methods

Normalization method	Calculation of normalized weight
Raw weights	Weight
Pointwise mutual information (PMI)	$\log_{10}\left(\frac{weight * ndocs}{w_1 df * w_2 df}\right)$

Sum of IDFs	$(w_1 idf + w_2 idf) * weight$
Product of IDFs	$(w_1 idf * w_2 idf) * weight$
Sum of document frequencies	$(w_1 df + w_2 df) * weight$
Product of document frequencies	$(w_1 df * w_2 if) * weight$
Inverse of sum of IDFs	$\frac{weight}{(w_1 idf + w_2 idf)}$
Inverse of prod of IDFs	$\frac{weight}{(w_1 idf * w_2 idf)}$
Inverse of sum of document frequencies	$\frac{weight}{(w_1 df + w_2 df)}$
Inverse of product of document frequencies	$\frac{weight}{(w_1 df * w_2 df)}$
Sum of frequencies	$(w_1 f + w_2 f) * weight$
Sum of frequencies multiplied by log sum of frequencies	$(w_1 f + w_2 f) * \log_{10}(w_1 f + w_2 f)$
Product of frequencies multiplied by log product of frequencies	$(w_1 f * w_2 f) * \log_{10}(w_1 f * w_2 f)$
Sum of frequencies divided by log sum of frequencies	$\frac{(w_1 f + w_2 f)}{\log_{10}(w_1 f + w_2 f)}$
Product of frequencies divided by log product of frequencies	$\frac{(w_1 f * w_2 f)}{\log_{10}(w_1 f * w_2 f)}$

Appendix B. ERP Participant assessment results

Table 9. ERP participant assessment results

Assessment	**Mean**	**SD**
Nelson-Denny Comprehension (raw score)	70.11	5.23
Nelson Denny reading rate (raw score)	298.47	94.35
PPVT (standard score)	119.74	10.56
TOWRE sight word (standard score)	103.53	9.63
TOWRE phonetic decoding (standard score)	101.37	9.90

Appendix C. ERP prediction performance

Table 10. Correlations between models and predictions, 20 iterations of 70/30 train/test.

Iteration	**Spearman**			**Pearson**		
	SVM-smGOLD	*SVM-LSA*	*LSA*	*SVM-smGOLD*	*SVM-LSA*	*LSA*
1	0.314	-0.044	0.069	0.304	-0.016	0.152
2	0.349	0.182	0.188	0.326	0.187	0.177
3	0.235	0.044	0.044	0.233	0.006	-0.001
4	0.335	0.054	0.078	0.323	0.069	0.118
5	0.246	0.007	0.007	0.218	-0.013	0.030
6	0.267	0.013	0.088	0.226	0.044	0.125
7	0.219	0.063	0.063	0.208	0.062	0.051
8	0.265	-0.020	0.115	0.242	-0.038	0.116
9	0.250	0.095	0.095	0.205	-0.026	0.036
10	0.192	0.106	0.106	0.147	0.013	0.039
11	0.150	0.079	0.079	0.140	0.038	0.045
12	0.233	0.154	0.154	0.223	0.117	0.108
13	0.200	-0.030	0.008	0.170	-0.052	0.016
14	0.238	0.054	0.054	0.215	0.084	0.082
15	0.129	0.094	0.094	0.133	0.056	0.056
16	0.357	0.092	0.092	0.300	0.022	0.011
17	0.175	-0.010	-0.010	0.195	-0.060	-0.080
18	0.009	0.026	0.027	-0.030	0.024	0.029
19	0.129	0.087	0.087	0.144	0.091	0.090
20	0.264	0.086	0.086	0.229	0.031	0.027
Min	0.009	-0.044	-0.010	-0.030	-0.060	-0.080
Max	0.357	0.182	0.188	0.326	0.187	0.177
Mean	0.228	0.057	0.076	0.208	0.032	0.061
SD	0.084	0.060	0.048	0.081	0.061	0.060

Appendix D. Stimuli for ratings and ERP study

Table 11. Stimuli and stimuli parameters for ratings and ERP.

Word1	**Word2**	**Category**	**Sim Rating**	**Assoc Rating**	**LSA 30**	**LSA 300**	**Word1 freq**	**W fre**
accuracy	case	random	1.45	2.70	0.80	0.02	1288	35
actress	bandage	random	1.16	1.21	-0.26	0.00	609	12
adultery	putty	random	1.39	1.09	-0.45	0.11	249	11
alpaca	cap	random	1.42	1.59	-0.25	0.06	151	28
apple	grape	Chiarello - similar	5.32	5.67	0.14	0.10	17029	48
army	navy	Chiarello - both	5.52	6.56	0.76	0.68	6615	16
artist	paint	Chiarello - associated	4.48	6.65	0.64	0.30	4579	39
assumption	rant	random	2.03	2.24	0.36	0.04	2729	19
assure	addition	random	1.39	1.21	0.36	0.22	1210	32
asylum	madhouse	Miller-Charles	5.97	5.94	0.07	0.03	471	19
atheism	pouch	random	1.00	1.12	-0.41	-0.06	6700	16
attractiveness	chili	random	1.26	1.15	-0.30	-0.06	296	68
authority	regime	random	4.71	4.53	0.82	0.24	3118	11
background	usage	random	1.55	1.85	0.71	0.12	6642	23
ball	bat	Chiarello - both	3.97	6.32	0.81	0.33	7764	19
banana	peach	Chiarello - similar	5.10	5.32	0.48	0.18	1586	36
barrel	council	random	1.35	1.15	0.33	0.00	2251	11
basin	sink	Chiarello - both	4.94	4.47	0.63	0.66	85	18
battle	director	random	1.55	1.62	0.78	0.17	4396	22
bear	twist	random	1.03	1.09	0.91	0.09	5815	32
bedroom	hypothesis	random	1.06	1.12	-0.28	0.01	2048	10
bee	honey	Chiarello - associated	4.45	6.88	0.51	0.35	799	26
bias	perception	random	4.39	4.94	0.93	0.51	3531	19
bigot	internship	random	1.16	1.26	-0.36	-0.08	429	50
birch	elm	Chiarello - similar	4.55	5.26	0.38	-0.16	76	11
bird	eagle	Thompson-Schill et al.	5.55	6.18	0.40	-0.03	2814	10
blackmail	protein	random	1.00	1.03	-0.38	-0.03	275	26
blanket	waste	random	1.19	1.18	0.17	0.01	1545	65
bloat	housemate	random	1.03	1.09	-0.40	-0.07	185	13
blouse	skirt	Chiarello - both	4.94	5.59	0.72	0.34	60	55
book	page	Chiarello - associated	4.94	6.45	0.78	0.12	26642	15
boy	clue	random	1.29	1.74	0.92	-0.01	9003	26
brand	pose	random	1.81	1.82	0.42	-0.05	5895	10
brandy	wine	Chiarello - both	5.33	5.74	0.51	0.20	83	31

Word1	Word2	Category	Sim Rating	Assoc Rating	LSA 30	LSA 300	Word1 freq	W fre
brass	iron	Chiarello - similar	5.23	5.06	0.78	0.14	663	34
brick	privacy	random	2.03	2.03	0.18	0.02	1410	27
bruise	stereotype	random	1.29	1.76	-0.33	0.00	223	16
brush	comb	Thompson-Schill et al.	5.74	6.62	0.46	0.20	1541	20
building	punishment	random	1.26	1.29	0.24	-0.04	10218	29
burlap	felt	Chiarello - similar	3.94	2.85	0.36	0.11	33	16
bus	mode	random	1.87	2.26	0.22	0.02	5125	42
butter	session	random	1.06	1.12	0.54	0.14	3434	15
bye	goodbye	Other	6.71	6.71	0.58	0.34	649	90
bystander	yeast	random	1.06	1.18	-0.32	0.00	152	90
camel	hump	Chiarello - associated	4.10	6.15	0.39	0.01	429	30
canada	steak	random	1.13	1.45	0.30	0.01	11553	19
candle	flame	Chiarello - associated	5.06	6.65	0.69	0.31	621	85
carbon	efficiency	random	2.06	4.12	0.81	0.60	1867	14
carrot	corn	Chiarello - similar	5.10	5.26	0.49	0.42	438	18
carry	executive	random	2.26	1.82	0.57	0.13	8404	12
casserole	gender	random	1.00	1.18	-0.40	-0.14	143	63
castle	designer	random	1.58	2.44	0.52	-0.01	1217	14
chapter	reason	random	1.35	1.65	-0.04	-0.01	1199	47
chip	penny	random	1.37	1.41	0.91	0.16	1783	14
church	theism	Other	4.00	3.65	0.52	0.81	11313	31
circle	cross	Chiarello - similar	2.65	3.24	0.67	0.34	3800	59
circus	clown	Chiarello - associated	4.45	6.65	0.57	0.24	464	85
clause	burden	random	1.87	1.47	0.81	0.06	1061	17
closet	vast	random	1.71	2.50	-0.09	-0.03	1535	39
cloth	dress	Chiarello - associated	5.10	5.39	0.60	0.18	614	37
cloud	output	random	1.48	1.18	0.90	0.28	2633	14
combination	animation	random	1.40	1.53	0.87	0.25	2555	14
companion	intuition	random	1.35	1.82	-0.04	0.10	694	34
compassion	brownie	random	1.48	2.00	-0.20	-0.02	870	18
complexity	porch	random	1.16	1.12	-0.43	-0.04	1067	63
concept	resource	random	2.55	2.85	0.73	0.22	7432	14
concert	lunch	random	1.35	1.79	0.82	0.07	1728	35
congressman	anime	random	1.19	1.00	-0.27	-0.09	355	41
consideration	tradition	random	1.87	1.74	0.85	0.09	1429	20
constitution	communism	random	1.94	3.47	0.84	0.30	3467	12
container	victim	random	1.32	1.47	-0.27	-0.04	1002	44
content	alternative	random	1.58	1.65	0.93	0.26	11623	39

Word1	Word2	Category	Sim Rating	Assoc Rating	LSA 30	LSA 300	Word1 freq	W fre
contrast	comparison	random	4.19	6.26	0.91	0.46	1557	47
cooker	commandment	random	1.23	1.09	-0.42	-0.01	264	12
correlation	coat	random	1.00	1.06	-0.35	0.00	1620	14
cotton	silk	Chiarello - similar	5.13	5.88	0.76	0.34	694	26
couch	philosophy	random	1.13	1.59	-0.42	-0.02	2423	59
cradle	baby	Chiarello - associated	4.13	5.88	0.30	0.05	227	14
crater	moon	Chiarello - associated	4.19	6.00	0.60	0.22	140	41
creationism	treadmill	random	1.03	1.00	-0.28	-0.01	678	51
crop	trigger	random	1.39	1.15	0.69	0.21	1115	25
cube	scroll	random	1.13	1.36	0.71	0.13	1409	14
currency	bolt	random	1.61	1.44	0.16	0.06	2240	10
custom	actor	random	1.58	1.88	0.10	-0.02	3001	27
cut	scissors	Thompson-Schill et al.	4.84	6.52	0.69	0.28	18614	48
decoy	duck	Chiarello - associated	2.19	1.97	0.43	0.04	120	28
deer	pony	Chiarello - similar	4.32	4.09	0.45	0.00	2169	79
definition	smell	random	1.45	1.18	-0.18	-0.04	7387	51
design	sweetheart	random	1.23	1.47	-0.26	-0.08	10014	24
desk	stool	Chiarello - similar	4.19	4.94	0.86	0.24	2955	22
devotion	milk	random	1.03	1.26	-0.13	-0.03	176	53
diaper	multiplier	random	1.26	1.21	-0.27	0.08	444	18
dirt	mud	Chiarello - both	6.32	6.70	0.85	0.45	1829	84
disagreement	tuna	random	1.00	1.09	-0.12	0.00	593	68
disgusting	gross	Other	6.35	6.82	0.78	0.56	3814	31
distinction	liar	random	1.52	2.24	0.02	0.11	1769	14
divorce	mother	random	2.58	4.24	0.93	0.67	1741	15
doom	agent	random	1.45	1.71	0.54	-0.02	1223	18
dorm	politics	random	1.23	1.79	-0.30	-0.04	807	91
dose	furniture	random	1.26	1.29	0.65	-0.05	1216	10
downstairs	jargon	random	1.06	1.24	-0.41	0.01	497	20
drums	piano	Chiarello - similar	4.58	5.68	0.68	0.67	966	12
ear	foot	Chiarello - similar	4.29	4.64	0.89	0.42	2783	56
elephant	paragraph	Other	1.00	1.21	0.45	-0.02	1227	16
empowerment	spaghetti	random	1.13	1.03	-0.25	-0.12	101	93
end	mess	random	1.45	1.65	0.94	0.09	47547	44
enforcement	net	random	1.83	2.29	0.82	-0.01	1792	37
engine	car	Chiarello - associated	4.61	6.32	0.38	0.20	4319	32
entry	score	random	2.87	2.94	0.78	0.29	2101	35
evidence	bead	random	1.26	1.12	-0.27	0.01	13829	10

Word1	Word2	Category	Sim Rating	Assoc Rating	LSA 30	LSA 300	Word1 freq	W fre
exam	gravity	random	1.32	1.82	0.19	0.02	1271	21
faith	shower	random	1.19	1.35	-0.40	-0.05	6813	34
farmer	plow	Chiarello - associated	3.81	5.88	0.21	-0.01	810	28
feature	tablet	random	1.58	2.33	0.93	0.67	4862	39
fever	obligation	random	1.32	1.26	-0.34	0.00	432	12
fiction	manager	random	1.13	1.21	0.11	-0.11	2755	45
fitness	vet	random	1.77	2.18	-0.19	-0.03	1953	16
flavor	tribe	random	1.29	1.38	-0.11	0.04	2038	56
flea	ant	Chiarello - similar	4.81	4.97	0.41	0.05	274	54
flew	regret	random	1.35	1.32	0.39	0.03	1121	29
fork	spoon	Thompson-Schill et al.	5.32	6.59	0.81	0.41	938	89
format	dispatcher	random	1.42	1.62	-0.45	-0.05	2238	10
fox	horse	Chiarello - similar	4.19	4.12	0.47	0.02	5528	49
freedom	beach	random	2.23	2.94	-0.13	-0.04	7782	23
frown	smile	Chiarello - both	3.65	6.00	0.41	0.51	178	38
gallon	jug	Chiarello - associated	4.68	5.71	0.77	0.62	1146	18
garage	piracy	random	1.58	1.41	-0.31	-0.06	1764	10
gas	lemonade	Other	1.16	1.32	0.65	0.17	8933	37
gaze	turtle	random	1.16	1.26	-0.01	0.08	219	95
gem	jewel	Miller-Charles	6.74	6.44	0.00	0.00	1504	11
gene	world	random	2.13	2.03	0.83	-0.02	1122	60
ghost	half	random	1.13	1.62	0.79	0.08	2307	27
grade	libertarian	random	1.23	1.76	-0.20	-0.04	7001	34
grammar	beauty	random	1.03	1.82	0.68	0.13	3164	23
grandson	query	random	1.29	1.50	-0.45	-0.04	220	24
graph	grandma	random	1.00	1.18	-0.16	-0.03	1231	22
grave	mileage	random	1.48	1.21	-0.31	-0.07	1058	68
grocer	store	Chiarello - associated	4.13	5.94	0.73	0.53	65	16
grumpy	grouchy	Other	6.55	6.53	0.56	0.34	754	34
guy	capitalist	random	2.06	1.97	-0.11	0.00	79747	14
habit	steam	random	1.10	1.06	0.67	-0.01	1841	54
hair	fur	Chiarello - similar	5.61	5.82	0.54	0.43	11644	88
happy	carpet	random	1.10	1.35	0.38	0.14	23716	11
harbor	boat	Chiarello - associated	3.87	5.88	0.65	0.16	514	37
hardware	section	random	1.77	3.03	0.65	-0.03	5085	49
head	leg	Chiarello - similar	4.10	5.24	0.94	0.32	27709	43
heckler	revenue	random	1.58	1.74	-0.29	-0.07	100	33
hermit	cave	Chiarello - associated	3.19	4.03	0.55	0.23	146	11

Word1	Word2	Category	Sim Rating	Assoc Rating	LSA 30	LSA 300	Word1 freq	W fre
hi	hello	Other	6.97	6.88	0.89	0.60	4112	29
hockey	ice	Chiarello - associated	3.81	6.59	0.70	0.21	4010	74
home	valley	random	1.90	2.35	0.53	-0.02	35632	10
house	lesson	random	1.61	2.18	0.66	-0.04	29295	26
hypocrisy	balance	random	1.29	1.68	0.45	-0.06	1042	51
ideology	razor	random	1.03	1.24	0.09	-0.01	1726	10
immigration	snow	random	1.03	1.29	0.09	0.03	1235	42
incident	destroy	random	2.39	2.61	0.21	0.16	2824	39
infection	treat	random	2.42	3.09	-0.14	0.22	1413	59
insight	blatant	random	1.94	1.74	0.56	0.12	1702	12
integer	buddy	random	1.26	1.15	-0.21	-0.04	202	47
involve	halfway	random	1.65	1.62	-0.07	-0.02	1765	16
jeep	plane	Chiarello - similar	3.81	4.09	0.79	0.29	523	37
jelly	jam	Chiarello - both	6.32	6.68	0.74	0.02	1254	13
jet	budget	random	1.19	2.15	0.51	0.24	1208	52
justification	eliminate	random	1.65	1.97	0.72	0.22	1421	13
justify	summer	random	1.13	1.03	-0.29	-0.19	3652	66
key	door	Chiarello - associated	3.90	6.29	0.14	0.18	7588	12
knock	warrant	random	2.10	2.85	0.19	0.08	2680	13
law	justice	Thompson-Schill et al.	5.32	6.50	0.87	0.35	24055	45
lawsuit	meaningless	random	1.45	2.06	0.45	0.02	1111	18
lawyer	nurse	Chiarello - similar	3.29	3.79	0.43	0.10	3001	18
layer	liquid	random	2.06	2.72	0.93	0.46	1677	26
leap	pen	random	1.32	1.18	0.53	0.06	1025	16
lee	grown	random	1.42	1.00	0.32	0.00	1420	36
legalization	toad	random	1.00	1.03	-0.47	-0.04	1142	14
lemon	pear	Chiarello - similar	4.68	5.00	0.56	0.20	1034	15
lie	sweet	random	1.06	1.47	0.24	0.01	7123	82
light	lamp	Thompson-Schill et al.	6.39	6.65	0.76	0.71	16912	72
lord	tab	random	1.23	1.03	0.08	0.04	3944	15
lotion	cream	Chiarello - both	5.90	6.12	0.74	0.31	355	36
machine	villain	random	1.45	1.76	0.22	0.06	8932	12
mad	anger	Thompson-Schill et al.	6.61	6.56	0.37	0.15	6534	23
man	woman	Chiarello - both	4.65	6.79	0.37	0.08	71832	22
management	chart	random	2.55	3.41	0.72	0.08	3810	13
market	carrier	random	2.13	2.53	0.81	0.08	16947	17
maximum	manufacturer	random	1.74	2.35	0.81	0.08	1620	11
meal	unfortunate	random	1.03	1.35	0.03	-0.17	3198	19

Word1	Word2	Category	Sim Rating	Assoc Rating	LSA 30	LSA 300	Word1 freq	W fre
medicine	amount	random	2.55	4.15	0.86	0.15	2674	20
met	texture	random	1.10	1.15	-0.24	-0.08	10417	10
miner	coal	Chiarello - associated	4.03	6.56	0.02	0.12	92	13
minimum	consumption	random	1.55	2.91	0.81	0.30	5352	17
minister	aroma	random	1.23	1.15	-0.04	0.01	1095	11
mint	candy	Chiarello - both	4.81	5.71	0.30	0.01	1129	30
mistaken	criticism	random	2.90	2.88	0.82	0.29	1620	27
modernism	wrist	random	1.29	1.15	-0.11	0.01	103	11
mold	bread	Chiarello - associated	3.03	4.75	0.66	0.31	652	38
mortgage	shown	random	1.32	1.47	0.34	-0.01	1245	43
moth	fly	Chiarello - both	5.52	5.50	0.49	0.19	273	48
mouse	rat	Chiarello - both	5.61	6.44	0.37	0.02	4177	11
movement	association	random	2.77	2.38	0.94	0.33	5406	12
mug	beer	Chiarello - associated	3.68	5.94	0.46	0.30	529	11
name	tortilla	random	1.06	1.18	-0.21	-0.01	34714	19
nationalist	cuddle	random	1.03	1.00	-0.45	-0.04	284	45
needle	thread	Thompson-Schill et al.	4.06	6.85	0.04	-0.14	819	18
needless	force	random	1.42	2.09	0.01	0.04	1403	14
nickel	dime	Chiarello - both	5.74	6.41	0.55	0.24	462	74
nightmare	tape	random	1.00	1.38	0.84	0.08	1679	29
onion	tears	Chiarello - associated	3.26	5.71	0.15	-0.01	1314	30
opinion	evening	random	1.16	1.44	-0.30	-0.12	19305	17
opportunity	contest	random	2.87	2.44	0.73	0.16	5308	14
orb	scum	random	1.39	1.29	-0.03	0.00	165	10
ounce	pound	Chiarello - both	4.84	6.24	0.74	0.47	623	19
outrage	deodorant	random	1.23	1.26	0.02	-0.07	876	22
oxygen	rating	random	1.35	1.53	0.41	-0.04	1251	12
paradox	valentine	random	1.26	1.24	-0.06	-0.04	816	36
patriarchy	raccoon	random	1.26	1.15	-0.48	-0.02	690	33
percentage	summary	random	2.39	2.38	0.62	0.22	3395	10
persuasion	seal	random	1.23	1.44	-0.07	-0.02	164	11
petty	attitude	random	3.19	3.76	0.90	0.37	1296	46
phenomenon	struggle	random	1.61	1.85	0.69	0.21	1232	22
pillow	fort	Other	2.57	4.62	0.46	0.14	920	60
platform	default	random	1.97	1.97	0.94	0.60	3735	45
poll	knife	random	1.26	1.24	0.12	-0.06	1440	43
pool	translate	random	1.13	1.26	0.08	-0.07	3752	14
pork	mentality	random	1.16	1.12	0.26	-0.01	1037	21

Word1	**Word2**	**Category**	**Sim Rating**	**Assoc Rating**	**LSA 30**	**LSA 300**	**Word1 freq**	**W fre**
prediction	diner	random	1.10	1.32	-0.21	0.00	807	22
pregnancy	glad	random	2.23	3.29	0.19	0.03	2269	13
press	pitch	random	1.93	1.76	0.84	0.14	5675	20
procreation	maple	random	1.26	1.06	0.04	-0.04	140	85
promote	identity	random	1.81	2.26	0.90	0.34	1881	25
prude	freezer	random	1.32	1.00	-0.43	-0.05	137	90
python	guilt	random	1.06	1.21	-0.01	0.04	3110	18
qualify	stable	random	1.90	2.12	0.61	0.15	1292	24
rage	farm	random	1.48	1.24	0.56	-0.03	3124	24
rake	leaf	Chiarello - associated	4.06	6.38	0.56	0.07	280	80
ram	edge	random	1.65	1.82	0.76	0.25	2294	39
raw	disagree	random	1.29	1.29	0.23	-0.03	2606	10
reassurance	pencil	random	1.03	1.15	-0.14	0.01	126	88
recommend	unity	random	1.35	1.91	0.76	0.22	7297	15
recover	sugar	random	1.39	1.47	0.48	0.32	1081	36
recovery	quest	random	2.45	2.21	0.63	0.12	1782	15
reform	apartment	random	1.32	1.76	-0.15	-0.09	1389	38
relativism	boxer	random	1.06	1.06	-0.20	-0.05	243	78
requirement	battery	random	2.00	2.61	0.55	-0.01	1537	57
retirement	task	random	1.77	1.85	0.38	0.07	1669	19
revolution	unknown	random	1.35	1.71	0.73	0.15	2402	17
righteousness	scan	random	1.23	1.24	0.00	-0.15	190	10
riot	procedure	random	1.90	1.62	0.54	0.03	1105	15
rob	require	random	1.42	1.29	0.08	-0.09	1338	64
robber	thief	Thompson-Schill et al.	6.26	6.85	0.84	0.19	238	85
rub	stream	random	1.42	1.29	0.10	0.03	1234	38
rubber	tire	Chiarello - associated	4.39	6.26	0.80	0.34	1350	13
rush	stuck	random	1.42	2.35	0.92	0.39	2755	77
salad	atheist	random	1.00	1.32	-0.38	-0.02	1077	71
scan	controller	random	2.19	2.50	0.70	0.05	1099	25
scenario	belief	random	2.07	2.12	0.68	-0.03	3354	56
school	apocalypse	random	1.16	1.68	0.25	-0.06	49862	10
script	eye	random	1.84	1.94	0.46	-0.03	2292	10
search	engineer	random	1.84	2.56	0.69	0.05	8026	34
sector	audio	random	1.84	1.59	0.42	0.02	1683	24
seem	hung	random	1.26	1.24	-0.21	-0.04	27491	15
semi	spin	random	1.45	1.76	0.63	0.27	3153	19
senate	safe	random	1.48	1.56	0.49	0.02	1367	10

Word1	Word2	Category	Sim Rating	Assoc Rating	LSA 30	LSA 300	Word1 freq	W fre
send	reflect	random	2.03	1.91	0.16	0.02	9255	14
sergeant	variety	random	1.16	1.18	-0.20	-0.01	323	27
set	role	random	2.87	3.15	0.14	0.01	26099	61
setup	menu	random	3.03	3.21	0.93	0.38	2495	27
shark	trout	Chiarello - similar	4.19	4.47	0.29	-0.04	1240	20
sheep	wool	Chiarello - associated	4.68	6.21	0.04	0.19	1392	34
shell	sea	Chiarello - associated	4.48	6.68	0.79	0.21	2350	33
shirt	polo	Other	5.35	6.00	0.60	0.36	5656	18
shoe	sandal	Other	5.50	5.82	0.44	0.18	1225	17
shoulder	chest	random	4.42	5.29	0.97	0.74	2584	32
sickness	health	Thompson-Schill et al.	3.74	6.29	0.46	0.30	395	12
skip	jump	Thompson-Schill et al.	5.19	5.88	0.70	0.25	2179	68
smoke	tobacco	Thompson-Schill et al.	4.81	6.85	0.30	0.42	5925	11
snake	mask	random	1.23	1.38	0.92	0.26	1796	20
socks	shoes	Other	4.94	6.65	0.82	0.63	1427	40
sofa	chair	Chiarello - both	5.58	5.85	0.71	0.46	263	26
sole	compliment	random	1.16	1.39	0.19	-0.05	1376	12
somebody	filter	random	1.10	1.15	0.32	0.05	8207	18
sort	license	random	1.16	1.38	0.25	0.03	22981	37
sound	union	random	1.84	1.74	0.13	0.02	20130	60
source	emotion	random	1.77	2.32	0.57	0.04	17256	16
speech	sin	random	1.42	1.65	0.86	0.14	6926	26
spider	web	Chiarello - associated	3.90	6.91	0.43	-0.01	2214	60
spirit	legacy	random	3.35	2.76	0.91	0.06	2548	10
stage	prize	random	2.55	3.56	0.66	0.12	4288	14
star	sky	Chiarello - associated	4.84	6.50	0.63	0.36	10093	34
station	trail	random	2.71	2.68	0.92	0.29	4406	12
stem	petal	Chiarello - similar	4.39	5.85	0.00	0.01	1373	19
sticker	monkey	random	1.23	1.32	0.38	0.01	1166	19
stigma	pint	random	1.16	1.09	-0.04	-0.10	906	44
stoop	avocado	random	1.00	1.03	-0.43	-0.10	162	26
stretch	cast	random	1.52	1.76	0.82	0.31	2278	35
string	rope	Chiarello - both	5.48	6.26	0.65	0.18	2527	11
sue	society	random	1.52	2.32	0.31	0.09	2198	15
sunflower	modesty	random	1.17	1.35	-0.41	0.03	113	12
surgery	equality	random	1.61	1.21	-0.33	-0.06	3665	25
symbol	suggestion	random	2.03	2.76	0.58	0.00	1421	18
syntax	broke	random	1.55	1.94	-0.23	-0.04	1008	70

Word1	Word2	Category	Sim Rating	Assoc Rating	LSA 30	LSA 300	Word1 freq	W fre
tack	nail	Chiarello - both	5.13	5.32	0.59	-0.04	247	18
team	immune	random	1.45	1.47	0.47	-0.01	20020	11
technology	heart	random	1.55	2.09	0.11	0.01	8367	10
teeth	camp	random	1.23	1.12	0.60	0.10	4559	29
text	prose	Other	4.29	4.31	0.63	0.30	8898	28
throw	toss	random	6.65	6.33	0.91	0.48	10469	17
tiger	lion	Chiarello - both	5.65	6.18	0.80	0.46	1335	17
till	slide	random	1.61	1.24	0.80	0.20	4137	21
tired	sleepy	Other	6.74	6.88	0.73	0.43	5570	29
tooth	react	random	1.35	1.62	-0.16	0.21	1105	22
tourist	dare	random	1.16	1.53	-0.11	0.05	801	26
tub	bath	Thompson-Schill et al.	6.19	6.74	0.87	0.81	872	12
tube	truth	random	1.06	1.15	-0.23	-0.08	1687	10
tulip	daisy	Chiarello - similar	5.61	6.12	0.15	-0.10	81	22
tuner	profession	random	1.84	1.74	-0.01	-0.03	120	10
twitter	audience	random	2.65	3.68	0.83	0.14	3628	42
typo	stranger	random	1.16	1.21	0.18	0.06	1053	21
tyranny	pepper	random	1.23	1.24	-0.34	-0.01	579	17
uncle	aunt	Chiarello - both	5.32	6.44	0.56	0.91	3232	16
unhappy	jerk	random	2.84	3.88	0.68	0.03	1024	33
uniform	weapon	random	2.52	3.74	0.72	0.27	1214	49
usher	movie	Chiarello - associated	2.32	3.32	0.28	0.20	122	33
velvet	linen	Chiarello - similar	4.19	4.91	0.56	0.25	193	66
verify	jury	random	3.52	3.79	0.44	0.20	1134	13
vermin	pan	random	1.39	1.09	-0.22	-0.02	110	20
wallpaper	daughter	random	1.29	1.38	-0.15	0.00	1087	61
wash	cook	random	3.35	4.68	0.73	0.38	2425	46
wave	ocean	Chiarello - associated	5.23	6.68	0.77	0.30	2198	23
way	immature	random	1.19	1.44	0.41	0.04	145795	12
weird	bud	random	1.26	1.38	0.63	0.15	16343	11
wife	instrument	random	1.26	1.35	0.03	-0.03	16363	11
winter	spring	random	4.97	6.24	0.92	0.57	4403	22
wolf	dog	Chiarello - both	5.42	5.38	0.48	0.77	1567	23
word	sentence	Other	4.42	6.41	0.80	0.65	24159	57
wrap	tournament	random	1.48	1.06	0.10	-0.08	1804	18
zone	gear	random	1.48	1.94	0.83	0.15	2812	47

Appendix E. Stimuli and stimuli parameters

Table 12. Word pairs from Chiarello et al. (1990)

Associated only		Similar and associated		Similar only	
alley	cat	ale	beer	apple	grape
apple	tree	arm	leg	arm	nose
artist	paint	army	navy	bacon	steak
bee	honey	ball	bat	banana	peach
bone	dog	basin	sink	bean	onion
book	page	blouse	skirt	bear	cow
button	coat	boot	shoe	birch	elm
camel	hump	brandy	wine	brass	iron
candle	flame	brush	comb	burlap	felt
cheese	mouse	butter	bread	car	ship
circus	clown	coat	hat	carrot	corn
cloth	dress	coffee	tea	circle	cross
cow	milk	cotton	wool	coat	gown
cradle	baby	dirt	mud	cotton	silk
crater	moon	doctor	nurse	dagger	rifle
crew	ship	dog	cat	deer	pony
crown	king	engine	motor	desk	stool
decoy	duck	figure	shape	drums	piano
engine	car	frown	smile	ear	foot
farmer	plow	inch	foot	flea	ant
fish	water	jacket	coat	floor	wall
flea	dog	jelly	jam	fox	horse
floor	wood	knife	fork	garlic	mint
gallon	jug	lizard	snake	gin	wine
grocer	store	lotion	cream	hair	fur
hammer	nail	man	woman	head	leg
harbor	boat	mint	candy	house	cabin
hermit	cave	moth	fly	jeep	plane
hockey	ice	mouse	rat	knife	pot
key	door	nickel	dime	lamp	chair
miner	coal	ounce	pound	lawyer	nurse
mold	bread	oven	stove	lemon	pear
mug	beer	pepper	salt	music	art
nest	bird	pot	pan	oak	maple
onion	tears	queen	king	orchid	tulip
pilot	plane	road	path	pan	bowl
rake	leaf	sea	ocean	pants	hat
rubber	tire	shirt	tie	roof	door
rug	floor	silver	gold	shark	trout
sheep	wool	sleet	snow	shoe	glove

Associated only		Similar and associated		Similar only	
shell	sea	sofa	chair	steel	brass
spider	web	steel	iron	stem	petal
star	sky	string	rope	street	path
stove	heat	sword	knife	sugar	salt
train	track	tack	nail	table	bed
usher	movie	tiger	lion	train	canoe
waist	belt	uncle	aunt	tulip	daisy
wave	ocean	wolf	dog	velvet	linen

Table 13. Word pairs from Plaut and Booth (2000).

Related		Unrelated	
adult	child	admit	learn
agony	pain	ahead	piece
alarm	clock	alike	post
argue	fight	allow	knee
birth	death	alone	death
blade	knife	anger	look
blank	empty	angle	tight
blaze	fire	apart	aunt
bored	tired	arrow	reef
bride	groom	avoid	talk
brief	short	basic	human
bring	take	beast	tree
canoe	boat	begin	open
chain	links	bench	tale
chuck	throw	blind	exit
cigar	smoke	bound	rain
clean	dirty	burst	yell
close	open	cabin	glue
coach	team	cause	south
coral	reef	charm	happy
court	judge	check	hotel
crane	lift	cheek	book
creek	river	chest	live
cycle	bike	chief	black
death	live	china	bird
ditch	hole	clear	music
donor	blood	climb	ghost
enter	exit	cloth	sharp
fairy	tale	cloud	watch
fence	post	color	year

Related		Unrelated	
flame	fire	count	bike
flood	water	crack	groom
fresh	fruit	crash	curse
funny	laugh	crawl	pain
ghoul	ghost	cream	fire
glove	hand	crowd	judge
grain	wheat	curve	move
grasp	hold	dense	fake
grass	green	dream	noise
heavy	light	drill	broom
honey	sweet	drink	dress
house	home	early	take
joint	knee	equal	treat
knock	door	event	green
labor	work	extra	call
large	small	faith	stop
lemon	lime	favor	fire
loose	tight	final	child
major	minor	floor	money
maple	tree	found	right
march	april	front	young
mint	candy	frost	bread
month	year	giant	smoke
motel	hotel	glory	decay
north	south	going	paper
novel	book	guard	knife
paint	brush	guest	steal
paste	glue	habit	plane
phone	call	hurry	laugh
phony	fake	leave	write
piano	play	level	door
pilot	plane	lower	short
poker	cards	meter	lion
print	write	model	turn
quack	duck	moist	throw
queen	king	motor	metal
radio	music	nerve	links
razor	sharp	never	work
reach	grab	notes	beach
scent	smell	nurse	path
shame	guilt	party	small
share	gives	patch	fruit
sheet	paper	pearl	duck
shift	gears	pitch	april
shirt	pants	plain	blood

Related		Unrelated	
shore	beach	plate	ocean
shout	yell	prize	sweet
skirt	dress	proud	bite
slice	piece	pupil	pants
smile	happy	quick	horse
snake	bite	raise	shoes
socks	shoes	rapid	fork
sound	noise	ready	light
spare	tire	reply	play
speak	talk	rifle	chair
spend	money	rough	lime
spoon	fork	scale	track
stall	horse	score	hold
stare	look	screw	clock
steel	metal	shape	home
still	move	shine	minor
stone	rock	shock	king
storm	rain	shoot	team
stuff	things	sight	hand
super	great	solid	brush
swear	curse	split	fight
sweep	broom	stalk	cards
table	chair	stamp	rock
teach	learn	stand	thing
thief	steal	state	great
tiger	lion	steam	candy
toast	bread	stiff	smell
tooth	decay	store	tire
touch	feel	straw	hole
trail	path	swamp	wheel
train	track	swift	guilt
trick	treat	tense	gear
truce	peace	today	water
twist	turn	topic	lift
wagon	wheel	total	peace
waves	ocean	tower	boat
white	black	trunk	tired
wings	bird	unite	dirty
wrist	watch	usual	river
wrong	right	visit	feel
youth	young	voice	give
		width	wheat
		worse	grab

References

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Pasca, M., & Soroa, A. (2009). A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. In *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL* (pp. 19–27).

Audet, C., & Burgess, C. (1999). Using a high-dimensional memory model to evaluate the properties of abstract and concrete words. In *Proceedings of the Cognitive Science Society* (pp. 37–42). Mahwah, NJ: Laurence Erlbaum Aasociates.

Barsalou, L.W., Santos, A., Simmons, W. K., & Wilson, C. D. (2008). Language and simulation in conceptual processing. In M. De Vega, A. M. Glenberg, & A. C. Graesser (Eds.), *Symbols, embodiment, and meaning* (pp. 245–283). Oxford: Oxford University Press.

Barsalou, Lawrence W. (1987). The instability of graded structure: implications for the nature of concepts. In U. Neisser (Ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization* (pp. 101–140). Cambridg.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi : An Open Source Software for Exploring and Manipulating Networks. In *International AAAI Conference on Weblogs and Social Media*.

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, *3*, 1137–1155.

Bird, S., Loper, E., & Klein, E. (2009). *Natural Language Processing with Python*. O'Reilly Media.

Blouw, P., & Eliasmith, C. (2003). A Neurally Plausible Encoding of Word Order Information into a Semantic Vector Space, 1905–1910.

Bolger, D. J., Balass, M., Landen, E., & Perfetti, C. (2008a). Context Variation and Definitions in Learning the Meanings of Words: An Instance-Based Learning Approach. *Discourse Processes*, *45*(2), 122–159. doi:10.1080/01638530701792826

Bolger, D. J., Balass, M., Landen, E., & Perfetti, C. a. (2008b). Context Variation and Definitions in Learning the Meanings of Words: An Instance-Based Learning Approach. *Discourse Processes*, *45*(2), 122–159. doi:10.1080/01638530701792826

Bolger, D. J., & Jackson, A. F. (n.d.). Acquiring Word Meaning: How are Contexts different from Definitions? *Under review*.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Budanitsky, A., & Hirst, G. (2006). Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, *32*(1), 14–47.

Bullinaria, J. a, & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: a computational study. *Behavior research methods*, *39*(3), 510–26. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17958162

Bullinaria, J. a, & Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD. *Behavior research methods*, *44*(3), 890–907. doi:10.3758/s13428-011-0183-8

Burgess, C. (2000). Theory and Operational Definitions in Computational Memory Models: A Response to Glenberg and Robertson. *Journal of Memory and Language*, *43*(3), 402–408. doi:10.1006/jmla.2000.2715

Burgess, C., & Livesay, K. (1998). The effect of corpus size in predicting reaction time in a basic word recognition task: Moving on from Kucera and Francis. *Behavior Research Methods, Instruments, & Computers*, *30*(2), 272–277.

Burgess, C., & Lund, K. (1997). Modelling Parsing Constraints with High-dimensional Context Space. *Language and Cognitive Processes*, *12*(2), 177–210. doi:10.1080/016909697386844

Burgess, C., & Lund, K. (1998). The Dynamics of Meaning in Memory. In Dietrich & Markman (Eds.), *Cognitive Dynamics: Conceptual Change in Humans and Machines* (pp. 1–23).

Burrows, S., & Tahaghoghi, S. M. M. (2007). Source code authorship attribution using n-grams. In *Proceedings of the 12th Australasian Document Computing Symposium* (pp. 32–39).

Butts, C. T. (2008). Social network analysis: A methodological introduction. *Asian Journal Of Social Psychology*, *11*(1), 13–41. doi:10.1111/j.1467-839X.2007.00241.x

Carlson, T. A., Simmons, R. A., Kriegeskorte, N., & Slevc, L. R. (2014). The Emergence of Semantic Meaning in the Ventral Temporal Pathway. *Journal of Cognitive Neuroscience*, *X*(Y), 1–12. doi:10.1162/jocn

Chelba, C., Bikel, D., Shugrina, M., Nguyen, P., & Kumar, S. (2012). *Large Scale Language Modeling in Automatic Speech Recognition* (pp. 1–6).

Cheng, B., & Titterington, D. M. (1994). Neural networks: A review from a statistical perspective. *Statistical Science*, *9*(1), 2–54.

Chiarello, C., Burgess, C., & Richards, L. (1990). Semantic and Associative Priming in the Cerebral Hemispheres : Some Words Do , Some Words Don ' t . . . Sometimes , Some Places, *104*, 75–104.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428.

Collins-Thompson, K., & Callan, J. (2007). Automatic and human scoring of word definition responses. In *Proceedings of the NAACL-HLT 2007 Conference* (pp. 476–483). Rochester.

Daalen-kapteijns, M. Van, & Elshout-mohr, M. (2001). Deriving the Meaning of Unknown Words From Multiple Contexts, (March), 145–181.

Dean, J., Corrado, G. S., Monga, R., Chen, K., Devin, M., Le, Q. V, … Ng, A. Y. (2012). Large Scale Distributed Deep Networks. In *Neural Information Processing Systems* (pp. 1–11).

Demsar, J., Curk, T., Erjavec, A., Gorup, C., Hocevar, T., Milutinovic, M., … Zupan, B. (2013). Orange: Data mining toolbox in Python. *Journal of Machine Learning Research*, *14*, 2349–2353.

Deyne, S. D. E., & Storms, G. (2008). Word associations : Network and semantic properties. *Behavior Research Methods*, *40*(1), 213–231. doi:10.3758/BRM.

Dijkstra, E. W. (1959). A Note on Two Problems in Connexion with Graphs. *Numerische Mathematick*, *1*, 269–271.

Dorogovtsev, S. N., & Mendes, J. F. (2001). Language as an evolving word web. *Proceedings. Biological sciences / The Royal Society*, *268*(1485), 2603–6. doi:10.1098/rspb.2001.1824

Dumais, S., Banko, M., Brill, E., Lin, J., & Ng, A. (2002). Web Question Answering: Is More Always Better ? In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*.

Durda, K., Caron, R., & Buchanan, L. (2010). An Application of Operational Research to Computational Linguistics : Word Ambiguity. *Information Systems and Operational Research*, *48*, 1–21.

Eck, N. J. Van, & Waltman, L. (2009). How to Normalize Co-Occurrence Data ? An Analysis of Some Well-Known Similarity Measures Nees Jan van Eck and Ludo Waltman REPORT SERIES.

Eifrem, E. (2009). Neo4j - The Benefits of Graph Databases. In *Qcon San Francisco*.

Federmeier, K. D., & Kutas, M. (1999). A Rose by Any Other Name: Long-Term Memory Structure and Sentence Processing. *Journal of Memory and Language*, *41*(4), 469–495. doi:10.1006/jmla.1999.2660

Finn, P. J. (1977). Word frequency, information theory, and cloze performance: A transfer feature theory of processing in reading. *Reading Research Quarterly*, *13*(4), 508–537.

Frishkoff, G. a, Collins-Thompson, K., Perfetti, C. a, & Callan, J. (2008). Measuring incremental changes in word knowledge: experimental validation and implications for learning and assessment. *Behavior research methods*, *40*(4), 907–25. doi:10.3758/BRM.40.4.907

Frishkoff, G. A., Perfetti, C. A., & Collins-Thompson, K. (2010). Lexical quality in the brain : ERP evidence for robust word learning from context. *Developmental Neuropsychology*, *35*(4), 376–403. doi:10.1080/875656412010480915

Fukkink, R. G., Blok, H., & de Glopper, K. (2001). Deriving Word Meaning from Written Context: A Multicomponential Skill. *Language Learning*, *51*(3), 477–496. doi:10.1111/0023-8333.00162

Gabrilovich, E., & Markovitch, S. (2007). Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *International Joint Conferences on Artificial Intelligence* (pp. 1606–1611).

Goldman, S. R., Hogaboam, T. W., Bell, L. C., & Perfetti, C. A. (1980). Short-Term Retention of Discourse During Reading. *Journal of Educational Psychology*, *72*(5), 647–655.

Green, A. E., Kraemer, D. J. M., Fugelsang, J. a, Gray, J. R., & Dunbar, K. N. (2010). Connecting long distance: semantic distance in analogical reasoning modulates frontopolar cortex activity. *Cerebral cortex (New York, N.Y. : 1991)*, *20*(1), 70–6. doi:10.1093/cercor/bhp081

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological review*, *114*(2), 211–44. doi:10.1037/0033-295X.114.2.211

Groppe, D. M., Urbach, T. P., & Kutas, M. (2011). Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. *Psychophysiology*, *48*(12), 1711–1725. doi:10.1111/j.1469-8986.2011.01273.x

Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

Halgren, E., Dhond, R. P., Christensen, N., Van Petten, C., Marinkovic, K., Lewine, J. D., & Dale, A. M. (2002). N400-like Magnetoencephalography Responses Modulated by Semantic Context, Word Frequency, and Lexical Class in Sentences. *NeuroImage*, *17*(3), 1101–1116. doi:10.1006/nimg.2002.1268

Hall, M. A. (1999). *Correlation-based Feature Selection for Machine Learning*. The University of Waikato.

Hartuv, E., & Shamir, R. (1999). A Clustering Algorithm based on Graph Connectivity. *Information Processing Letters*, *76*(4), 1–9.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: what is it, who has it, and how did it evolve? *Science*, *298*(5598), 1569–79. doi:10.1126/science.298.5598.1569

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers*, *16*, 96–101.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, *79*, 2554–2558.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*, 359–366.

Howard, M. W., Addis, K. M., Jing, B., & Kahana, M. J. (2005). Semantic structure and episodic memory. In T. K. Landauer, D. Mcnamara, S. Dennis, & W. Kintsch (Eds.), *Latent Semantic Analysis : A Road to Meaning*. La uren ce E r lbau m.

Hughes, T., & Ramage, D. (2007). Lexical Semantic Relatedness with Random Graph Walks, (June), 581–589.

Hutchison, K. A. (2003). Is semantic priming due to association strength or feature overlap ? A micro analytic review. *Psychonomic Bulletin & Review*, *10*(4), 785–813.

Inkpen, D. (2007). A statistical model for near-synonym choice. *ACM Transactions on Speech and Language Processing*, *4*(1), 1–17. doi:10.1145/1187415.1187417

Islam, A., & Inkpen, D. (2008). Semantic text similarity using corpus-based word similarity and string similarity. *ACM Transactions on Knowledge Discovery from Data*, *2*(2), 1–25. doi:10.1145/1376815.1376819

Jackson, A. F., & Bolger, D. J. (n.d.). Neurophysiological Markers of Learning Word Meaning from Context. *In preparation*.

Jarmasz, M. (2003). *Roget's thesaurus as a lexical resource for natural language processing*.

Jenkins, J. R., Stein, M. L., & Wysocki, K. (1984). Learning vocabulary through reading. *American Educational Research Journal*, *21*(4), 767–787.

Jiang, J. J., & Conrath, D. W. (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *Proceedings of International Conference Research on Computational Linguistics*.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological review*, *114*(1), 1–37. doi:10.1037/0033-295X.114.1.1

Kaan, E. (2007). Event-Related Potentials and Language Processing: A Brief Overview. *Language and Linguistics Compass*, *1*(6), 571–591. doi:10.1111/j.1749-818X.2007.00037.x

Kakkonen, T., Myller, N., & Sutinen, E. (2006). Applying part-of-speech enhanced LSA to automatic essay grading. In *Proceedings of the 4th IEEE International Conference on Information Technology: Research and Education* (pp. 500–504).

Kakkonen, T., Myller, N., Timonen, J., & Sutinen, E. (2005). Automatic Essay Grading with Probabilistic Latent Semantic Analysis. In *Proceedings fo the 2nd Workshop on Building Educational Applications using NLP* (pp. 29–36).

Kintsch, W., & Mangalath, P. (2011). The Construction of Meaning. *Topics in Cognitive Science*, *3*(2), 346–370. doi:10.1111/j.1756-8765.2010.01107.x

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*(5), 363–394.

Koivisto, M., & Revonsuo, A. (2001). Cognitive representations underlying the N400 priming effect. *Cognitive brain research*, *12*(3), 487–90. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11689310

Kolb, P. (2006). Experiments on the difference between semantic similarity and relatedness. In *Proceedings of the 17th Nordic Conference of Computational Linguistics* (pp. 81–88).

Kutas, M., & Federmeier, K. D. (2011). Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Reviews Psychology*, *62*, 14.1–14.27. doi:10.1146/annurev.psych.093008.131123

Kwantes, P. J. (2005). Using context to build semantics. *Psychonomic Bulletin & Review*, *12*(4), 703–10. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16447385

Landauer, T. K., & Dumais, S. T. (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, *104*(2), 211–240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2-3), 259–284. doi:10.1080/01638539809545028

Landauer, T. K., Laham, D., & Foltz, P. (1997). Learning Human-like Knowledge by Singular Value Decomposition: A Progress Report. In *Proceedings of the 1997 Conference on Advances in Neural Information Processing Systems*.

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics : (de)constructing the N400. *Nature Review of Neuroscience*, *9*, 920–933. doi:10.1038/nrn2532

Ledoux, K., Camblin, C. C., Swaab, T. Y., & Gordon, P. C. (2006). Reading words in discourse: The modulation of lexical priming effects by message-level context. *Behavioral and Cognitive Neuroscience Reviews*, *5*(3), 107–127.

Lee, S., Baker, J., Song, J., & Wetherbe, J. C. (2010). An Empirical Comparison of Four Text Mining Methods. In *43rd Hawaii International Conference on System Sciences* (pp. 1–10). Ieee. doi:10.1109/HICSS.2010.48

Levin, E., Sharifi, M., & Ball, J. (2006). Evaluation of Utility of LSA for Word Sense Discrimination. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL* (pp. 77–80).

Li, C., Sun, A., & Datta, A. (2011). A Generalized Method for Word Sense Disambiguation based on Wikipedia. In *ECIR'11 Proceedings of the 33rd European Conference on Advances in Information Retrieval*.

Lin, D., & Pantel, P. (2002). Concept Discovery from Text. In *Proceedings of ACM Special Interest Group on Information Retrieval* (pp. 199–206). Tampere, Finland.

Lotte, F., Congedo, M., Lécuyer, a, Lamarche, F., & Arnaldi, B. (2007). A review of classification algorithms for EEG-based brain-computer interfaces. *Journal of neural engineering*, *4*(2), R1–R13. doi:10.1088/1741-2560/4/2/R01

Lowe, W. (2000). What is the Dimensionality of Human Semantic Space ? In *Proceedings of the 6th Neural Computation and Psychology Workshop* (pp. 303–311).

Luke, D. a, & Harris, J. K. (2007). Network analysis in public health: history, methods, and applications. *Annual review of public health*, *28*, 69–93. doi:10.1146/annurev.publhealth.28.021406.144132

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203–208.

Lund, K., Burgess, C., & Atchley, R. A. (1995). Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the Cognitive Science Society* (pp. 660–665). Hillsdale, N.J.: Erlbaum Publishers.

Magliano, J. P., & Graesser, A. C. (2012). Computer-based assessment of student-constructed responses. *Behavior research methods*, *44*(3), 608–21. doi:10.3758/s13428-012-0211-3

Marton, Y., Mohammad, S., & Resnik, P. (2009). Estimating Semantic Distance Using Soft Semantic Constraints in Knowledge-Source – Corpus Hybrid Models. In *Conference on Empirical Methods in Natural Language Processing*.

Mcdonald, S., & Ramscar, M. (2000). Testing the Distributional Hypothesis : The Influence of Context on Judgements of Semantic Similarity Distributional Models of Word Meaning. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*.

Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology*, *32*(341), 89–115.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Team, T. G. B., … Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176–182. doi:10.1126/science.1199644.Quantitative

Mihalcea, R., Corley, C., & Strapparava, C. (2005). Corpus-based and Knowledge-based Measures of Text Semantic Similarity. In *Proceedings of the 21st National Conference on Artificial Intelligence* (pp. 775–780).

Miller, G. a. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39–41. doi:10.1145/219717.219748

Minkov, E., & Cohen, W. W. (2008). Learning Graph Walk Based Similarity Measures for Parsed Text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 907–916).

Osterhout, L., Kim, A., & Kuperberg, G. (2006). The Neurobiology of Sentence Comprehension. In *The Cambridge Handbook of Psycholinguistics* (pp. 1–23). Cambridge: Cambridge University Press.

Padó, S., & Lapata, M. (2006). Dependency-based Construction of Semantic Space Models. *Computational Linguistics*, *33*(2), 161–199.

Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, *435*(7043), 814–8. doi:10.1038/nature03607

Parviz, M., Johnson, M., Johnson, B., & Brock, J. (2011). Using Language Models and Latent Semantic Analysis to Characterise the N400m Neural Response. In *Proceedings of Australasian Language Technology Association Workshop* (pp. 38–46).

Perfetti, C. A., Wlotko, E. W., & Hart, L. A. (2005). Word learning.pdf. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(6), 1281–1292.

Piattelli-Palmarini, M. (1980). How hard is the "hard core" of a scientific program? In *Language and Learning: The Debate between Jean Piaget and Noam Chomsky (the Royaumont debate)* (pp. 1–20).

Plaut, D. C., & Booth, J. R. (2000). Individual and developmental differences in semantic priming: empirical and computational support for a single-mechanism account of lexical processing. *Psychological review*, *107*(4), 786–823. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11089407

Radev, D., & Mihalcea, R. (2008). Networks and Natural Language Processing. *Artificial Intelligence Magazine*, *29*(3), 16–28.

Recchia, G., & Jones, M. N. (2009). More data trumps smarter algorithms : Comparing pointwise mutual information with latent semantic analysis. *Behavior Research Methods*. doi:10.3758/BRM.More

Rehurek, R., & Sojka, P. (2004). Software Framework for Topic Modelling with Large Corpora. In *LREC 2010*.

Rodd, J. M., Gaskell, M. G., & Marslen-Wilson, W. D. (2004). Modelling the effects of semantic ambiguity in word recognition. *Cognitive Science, 28*, 89–104. doi:10.1016/j.cogsci.2003.08.002

Rogers, T. T., & McClelland, J. L. (2011). Semantics without categorization. In E. M. Pothos & A. J. Wills (Eds.), *Formal Approaches in Categorization* (pp. 88–119). Cambridge: Cambridge University Press.

Rohde, D. L. T., Gonnerman, L. M., & Plaut, D. C. (2005). An Improved model of semantic similarity based on lexical co-occurrence. *Unpublished manuscript*.

Rubenstein, H., & Goodenough, J. B. (1965). Contextual Correlates of Synonymy. *Computational Linguistics*, *8*(10), 627–633.

Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to Student's t-test and the Mann-Whitney U test. *Behavioral Ecology*, *17*(4), 688–690. doi:10.1093/beheco/ark016

Schaeffer, S. E. (2007). Graph clustering. *Computer Science Review*, *1*(1), 27–64. doi:10.1016/j.cosrev.2007.05.001

Schatz, E. K., & Baldwin, R. S. (1986). Context clues are unreliable predictors of word meanings. *Reading Research Quarterly*, *21*(4), 439–453.

Scheel, S. L. (1998). *French language purism: French linguistic development and current national attitudes*.

Schutze, H. (Xerox P. A. R. C. (1998). Automatic Word Sense Discrimination. *Computational Linguistics*, *24*(1), 97–123.

Sereno, S. C., Brewer, C. C., & O'Donnell, P. J. (2003). Context effects in word recognition: evidence for early interactive processing. *Psychological science : a journal of the American Psychological Society / APS*, *14*(4), 328–33. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12807405

Shaoul, C., & Westbury, C. (2010). Exploring lexical co-occurrence space using HiDEx. *Behavior research methods*, *42*(2), 393–413. doi:10.3758/BRM.42.2.393

Silva, T. C., & Amancio, D. R. (2013). Discriminating word senses with tourist walks in complex networks. *The European Physical Journal B*, *86*(7), 297. doi:10.1140/epjb/e2013-40025-4

Silver, R. A. (2010). Neuronal arithmetic. *Nature Reviews Neuroscience*, *11*(7), 474–489. doi:10.1038/nrn2864

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cognitive science*, *29*(1), 41–78. doi:10.1207/s15516709cog2901_3

Strube, M., & Ponzetto, S. P. (2006). WikiRelate! Computing Semantic Relatedness Using Wikipedia. In *AAAI'06 Proceedings of the 21st National Conference on Artificial intelligence* (pp. 1419–1424).

Sun, Z., Wang, H., Wang, H., Shao, B., & Li, J. (2012). Efficient Subgraph Matching on Billion Node Graphs. In *Proceedings of the VLDB Endowment* (pp. 788–799).

Swanborn, M. S. L., & de Glopper, K. (1999). Incidental Word Learning while Reading : A Meta-Analysis. *Review of Educational Research*, *69*(3), 261–285.

Swanborn, M. S. L., & de Glopper, K. (2002). Impact of Reading Purpose on Incidental Word Learning From Context. *Language Learning*, *52*(1), 95–117. doi:10.1111/1467-9922.00178

Thompson-schill, S. L., Kurtz, K. J., & Gabrieli, J. D. E. (1998). Effects of Semantic and Associative Relatedness on Automatic Priming, *458*(38), 440–458.

Thornhill, D. E., & Van Petten, C. (2012). Lexical versus conceptual anticipation during sentence processing: Frontal positivity and N400 ERP components. *International journal of psychophysiology : official journal of the International Organization of Psychophysiology*. doi:10.1016/j.ijpsycho.2011.12.007

Tivarus, M. E., Ibinson, J. W., Hillier, A., Schmalbrock, P., & Beversdorf, D. Q. (2006). An fMRI study of semantic priming: modulation of brain activity by varying semantic distances. *Cognitive and behavioral neurology : official journal of the Society for Behavioral and Cognitive Neurology*, *19*(4), 194–201. doi:10.1097/01.wnn.0000213913.87642.74

Tremblay, A., & Newman, A. J. (2013). Modelling Non-linear Relationships in ERP Data Using Mixed-effects Regression with R Examples.

Tsang, V., & Stevenson, S. (2010). A Graph-Theoretic Framework for Semantic Distance, (December 2007).

Turney, P. D. (2001). Mining the Web for Synonyms : PMI-IR versus LSA on TOEFL PMI-IR. In *Proceedings of the 12th European Conference on Machine Learning* (pp. 1–12).

Utsumi, A. (2010). Exploring the Relationship between Semantic Spaces and Semantic Relations. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (pp. 257–262).

Van Daalen-Kapteijns, M., & Elshout-Mohr, M. (1981). The Acquisition of Word Meanings as a Cognitive Learning Process. *Journal Of Verbal Learning And Verbal Behavior*, *20*, 386–399.

Van Daalen-Kapteijns, M., Elshout-mohr, M., & de Glopper, K. (2001). Deriving the Meaning of Unknown Words From Multiple Contexts. *Language Learning*, *51*(1), 145–181.

Velik, R. (2008). Discrete Fourier Transform Computation Using Neural Networks. *2008 International Conference on Computational Intelligence and Security*, 120–123. doi:10.1109/CIS.2008.36

Wang, T., & Hirst, G. (2010). Near-synonym Lexical Choice in Latent Semantic Space. In *Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 1182–1190).

Weeds, J., & Weir, D. (2005). Co-occurrence Retrieval : A Flexible Framework for Lexical Distributional Similarity. *Computational Linguistics*, *31*(4), 439–475.

Weeds, J., Weir, D., & McCarthy, D. (2004). Characterising Measures of Lexical Distributional Similarity. In *Proceedings of the 20th international conference on Computational Linguistics*.

Weinberger, K. Q., Blitzer, J., & Saul, L. K. (2009). Distance Metric Learning for Large Margin Nearest Neighbor Classification. *The Journal of Machine Learning Research*, *10*, 207–244.

Widdows, D., & Dorow, B. (2002). A Graph Model for Unsupervised Lexical Acquisition. In *Proceedings of the 19th international conference on Computational linguistics*.

Wiemer-Hastings, P. (2000). Adding syntactic information to LSA. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*.

Xu, P., & Jelinek, F. (2004). Random forests in language modeling. In *Empirical Methods on Natural Language Processing*.