Spring 5-15-2021

Modeling Semantic Structure and Spreading Activation in Retrieval Tasks

Abhilasha Ashok Kumar
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Department of Psychological & Brain Sciences

Dissertation Examination Committee:
David A. Balota, Chair
Ian Dobbins
Jan Duchek
Brett Hyde
Jeffrey M. Zacks

Modeling Semantic Structure and Spreading Activation in Retrieval Tasks

by
Abhilasha A. Kumar

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2021
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgements

I thank my advisor, Dave Balota, for his continued mentorship and guidance throughout my graduate career. Dave's scientific rigor and curiosity has taught me invaluable lessons in research and also led me to discover new interests along the way. I thank him for motivating me to ask the right questions, encouraging me to discover my own niche in the field, and supporting my professional journey.

I am grateful to Jeff Zacks for always being a great source of advice and brilliant insights on research and professional matters, and for his helpful comments on earlier drafts of this dissertation. I also thank Ian Dobbins, whose thoughtful feedback on research methodology has always enriched my work, and especially this dissertation. I thank Jan Duchek for her constant support and encouragement, her insightful and constructive comments on my work throughout graduate school, and for showing me what excellent teaching should look like in a classroom. I also thank Brett Hyde for serving on my dissertation committee and motivating me to think about this dissertation from multiple angles.

I am thankful to Simon De Deyne and Adrian Staub for sharing the data from the Small World of Words project and the Cloze task, respectively, which has made this dissertation possible. I thank Armand Rotaru for his insightful thoughts on process modeling during early stages of this project. I am also extremely grateful to Mark Steyvers for nurturing my interest in computational modeling, which has motivated this dissertation and my general career trajectory.

I am grateful to my current and past lab members from the Cognitive Psychology Lab. I have enjoyed being part of this brilliant community of scientists and will cherish our interesting lab meetings (and snacks), conference meet-ups, and priming-related jokes. I especially want to thank Pete Millar, who patiently answered all my inane questions throughout graduate school

and set the bar high with his exceptional work ethic. I also thank Jessica Nicosia for helping me with several projects, piloting, and last-minute technical needs! I will miss standing outside our offices and chatting about all things WashU. I also thank Emily Cohen-Shikora and Andy Aschenbrenner for their insightful comments on several analyses, posters, papers, and talks! I am also grateful to our lab assistants (Tonya Waldeier and Cole Bolton) who were always available to help and somehow managed to keep the chaos to a minimum! Finally, I thank all the exceptional research assistants who have worked with me over the years (especially Emily Albertina, Michael Chang, Sabrina Hu, and Jasper Wilson) and have inspired me through their hard work and enthusiasm, and without whom no research would ever get done in the lab.

I am grateful to Mike Strube for reigniting my interest in statistics and modeling and teaching me how to teach by example. I am also thankful to all other instructors and faculty at WashU who have helped shape my research and teaching interests - Henry Biggs, Henry Chai, James Daily, Kristin Van Engen, Denise Head, John Nestojko, Heather Rice, Mitch Sommers, and Rebecca Treiman. I also thank David Archer for his invaluable support on technical matters throughout graduate school. I am also forever indebted to Ritwik Agarwal, Sudheendra Hangal, and Allyson Rosen for being my first mentors who steered me towards cognitive science.

I thank my friends who have been an invaluable support system during these past five years and made St. Louis a home for me - Marta, Meghan, Eylul, Matt, Kyle, Brent, Daphne, Benedict, and Jackson. I will cherish our many offline and online hangouts, barbecue and holiday parties, happy hour conversations, game nights, Diwali and Holi celebrations, and trips! I also thank Arushi, Ashwini, Ishita, and Sakshi for keeping me grounded and reassuring me that some friendships go beyond continents, professions, and pandemics.

I thank Pallavi and Ajay, for bringing Arudra into my life, for always being there in the scariest of times, and for being the glue that keeps our family together despite being countries (and continents) apart. I thank Papa, for being open to unlearning and for expecting nothing but the best from me, and Mom, for supporting my unconventional decisions and for instilling ambition and independence in me.

Finally, I thank my partner, Nick, for making me laugh so much, for being my best friend, and for forgetting the world with me.

Abhilasha A. Kumar

Washington University in St. Louis

May 2021

ABSTRACT OF THE DISSERTATION

Modeling Semantic Structure and Spreading Activation in Retrieval Tasks

by

Abhilasha A. Kumar

Doctor of Philosophy in Psychological & Brain Sciences

Washington University in St. Louis, 2021

Professor David A. Balota, Chair

Considerable work in the past decade has focused on representational accounts of how semantic information is acquired and organized, leading to the advent of modern Distributional Semantic Models (DSMs) that learn word meanings by extracting statistical information from large text corpora. However, mechanistic accounts for how meaning-related information is accessed and retrieved from semantic representations to ultimately produce responses within semantic tasks remain relatively understudied, especially for production-based tasks that require the selection of a single response amongst several activated competitors, such as in free association and sentence completion tasks. This dissertation evaluated the extent to which state-of-the-art DSMs combined with algorithmic and process models account for performance in two familiarity-driven tasks (relatedness and similarity judgments) and two production-based tasks (free association and sentence completion). Model comparisons revealed that while a process-based model based on the spreading activation mechanism successfully accounted for relatedness and similarity judgments, an interactive model based on word frequency and semantic similarity, combined with a thresholding function that incorporated competition from neighboring words best accounted for free association responses and response latencies. In addition, the results indicated that when participants produced multiple responses in the free association task, the second response was highly dependent upon the first response, instead of primarily being driven by the cue. In predicting Cloze sentence completion performance, a contextual “attention”-based DSM significantly outperformed other models, suggesting that information is accessed and retrieved in a syntactically constrained manner in language production tasks. Collectively, these findings shed light on how meaning-related information is activated and responses are differentially produced depending upon task demands. Importantly, there appears to be little evidence for a task-independent model of semantic memory representation, indicating the importance of incorporating both task-specific retrieval mechanisms and different representational formats in theories of semantic memory structure and processing. Abandoning a common semantic representation for models of knowledge-driven tasks is a major departure from previous approaches.

Chapter 1:
Introduction

Investigating the structure and organization of semantic memory has historically been at the forefront of explorations in cognitive psychology, natural language processing, and linguistics, due to its fundamental implications for understanding cognitive behavior and developing language-based tools and technologies. The last few decades have seen remarkable advances in explicitly modeling the structure of semantic memory. In particular, there has been an explosion of computational models of semantic memory that propose explicit mechanisms for how humans learn word meaning from natural language. These models, collectively called "distributional semantic models" (DSMs), are consistent with the "distributional hypothesis" (Firth, 1957; Harris, 1954), according to which words that occur in similar contexts tend to develop similar meanings. DSMs apply this intuition to large-scale text corpora (e.g., Wikipedia database, Google News articles, etc.), and construct semantic representations by applying statistical methods to infer which words occur in similar contexts. DSMs differ in how they define "context" (e.g., a window of words, sentences, or documents) and the core mechanism for learning representations (e.g., inferring latent dimensions, prediction, etc.; for a review, see Kumar, 2020). Typically, DSMs represent words in a high-dimensional space, where each concept is represented through a multidimensional vector and the angle between the vectors (or cosine similarity) within this high-dimensional space is indicative of semantic similarity between the concepts (see more detailed discussion of the specific DMSs tested below).

Collectively, DSMs have shown unprecedented success at explaining performance across different semantic tasks such as relatedness judgments (for a review, see Baroni, Dinu, &

Kruszewski, 2014; Günther, Rinaldi, & Marelli, 2019; Turney & Pantel, 2010), categorization (Lazaridou, Pham, & Baroni, 2015; Mikolov et al., 2013), and sentence comprehension (Devlin, Chang, Lee, & Toutanova, 2019). For example, modern DSMs trained on large text corpora and based on error-driven or error-free learning mechanisms such as word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2018) can successfully solve verbal analogy problems (e.g., king : man :: queen : ??), whereas more recent recurrent and "attention-based" neural networks such as BERT (Devlin et al., 2019; discussed in detail in Chapter 4) have made significant strides in modeling complex language tasks like question answering and coreference resolution. Although there has been considerable progress in explaining variance in these tasks, it is also worth noting that there is considerable variance left to be explained in terms of accounting for human baselines in semantic tasks (Ettinger, 2020; Niven & Kao, 2019).

Despite their success, there is a growing concern regarding the psychological plausibility of DSMs (Günther et al., 2019; Niven and Kao, 2019) and the extent to which these models mirror human cognition. Specifically, the semantic modeling enterprise has focused mostly on *representational* or *structural* accounts of semantic memory, i.e., how concepts are acquired, stored, and represented. For example, in a typical test of a DSM, the cosine similarity between concepts is used to predict whether two words are related or not (see Landauer & Dumais, 1997). However, *mechanistic* accounts that adequately explain the dynamics of retrieval from semantic memory within a particular task have not received the same kind of attention. Further, different models are not always evaluated on the same set of tasks, and it is unclear why their performance is better on tasks such as synonym detection (Bullinaria & Levy, 2007) and similarity judgments (Baroni et al., 2014) and worse when it comes to accounting for semantic priming effects (Hutchison, Balota, Cortese, & Watson, 2008; Mandera et al., 2017), free association performance (Griffiths, Steyvers, & Tenenbaum, 2007) and complex inference tasks (Niven & Kao, 2019). Therefore, there is a gap in the literature when it comes to *explaining* the mechanisms through which concepts are retrieved from semantic memory and how that may influence performance across different semantic tasks.

1.1 Not All Semantic Tasks Tap Similar Operations

An important aspect of studies that evaluate the predictive power of different DSMs is that most of this work has focused on similarity-type tasks (e.g., predicting similarity or relatedness judgments, lexical decision RTs, etc.), which primarily reflect familiarity or recognition-based processes (Balota & Chumbley, 1984). Specifically, these tasks may reflect overall activation within a memory network, whereas other tasks may demand the selection of a *single* word, which may lead to competition amongst activated representations. This may be particularly important in language production tasks, where multiple words may fit a given context, and competitors may need to be suppressed to produce a response. Indeed, this would suggest a distinction between *automatic* processes for similarity-based decisions and more *attention*-based processes involved in the selection of a single response (Neely, 1977)[1]. In this light, production-based language tasks may be ideal candidates for attention-based search of semantic memory. Unfortunately, relatively little work has attempted to model performance (in terms of responses and latencies) in these types of tasks using distributional semantic models. Therefore, investigating how distributional models of semantic memory account for **both** familiarity-based and production-based tasks is an important next step in the field.

[1] The automatic-attentional distinction has also been made in decision-making research (e.g., Kahneman, 2011), and distributional models have recently been applied to study heuristics, biases, and everyday decision-making (e.g., Singh, Richie, & Bhatia, 2020; Zou & Bhatia, 2019).

1.2 Spreading Activation and Free Association

One theoretical framework for understanding the mechanisms involved in retrieval from semantic memory is to conceptualize semantic memory as a large network, in line with early work by Collins and Quillian (1969). Within this network-based framework, activation is assumed to spread from one word to another, therefore providing insight into the temporal dynamics of word activation and retrieval (Collins & Loftus, 1975). The spreading activation mechanism has been posited to explain a wide variety of empirical phenomena including semantic priming effects in lexical decision (Neely, 2012), facilitation and disruption in lexical retrieval (Kumar & Balota, 2020), mediated priming in pronunciation (Balota & Lorch, 1986), long-distance priming effects in progressive demasking (Kumar, Balota, & Steyvers, 2019) and relatedness judgments (Kenett et al., 2017), and the dynamics of sentence production (Dell, 1986). Indeed, spreading activation has been a central retrieval mechanism in general models of cognition such as Anderson and Bower's (1973) Human Associative Memory (HAM) model, and Anderson's (1996) Adaptive Control of Thought (ACT) model.

In contrast to the distributional approach, which is typically based on large corpora of natural language, recent approaches within the semantic network tradition have attempted to quantitatively model semantic memory networks and spreading activation using responses from the free association task. Free association is a common task in semantic memory research, in which participants are asked to produce the first word (or words) that come to mind in response to a cue word. Nelson et al.'s (2004) University of South Florida (USF) free association norms (hereafter referred to as the USF norms), which consist of aggregate analyses of discrete responses collected for over 5,000 English words across several hundreds of participants, has historically been considered the gold standard of association in memory research (McRae, Khalkhali, & Hare, 2012), and has been cited over 2,000 times, based on Google Scholar citations. More recently, researchers have measured free association responses in a more continual manner, by asking participants to produce all responses within a certain time period (e.g., Kenett, Kenett, Ben-Jacob, & Faust, 2011) or a certain number of responses (e.g., three; De Deyne & Storms, 2008; De Deyne et al., 2019) to cue words. The responses from studies of free association have been used to construct large-scale semantic networks (e.g., De Deyne et al., 2019; Kenett et al., 2011; Steyvers & Tenenbaum, 2005), that have since been widely applied to several semantic tasks such as verbal fluency (Abbott, Austerweil, & Griffiths, 2015), episodic free recall (Kenett et al., 2017), and creative word association (Kenett et al., 2014), among others.

It is important to note that the use of free associations to create semantic networks and model memory structure is a controversial issue within the distributional modeling literature (Jones, Hills, & Todd, 2015). Specifically, using human-generated associations to explain human behavior in other semantic tasks may indicate shared variance between tasks, and may not represent a true account of semantic memory organization (but see De Deyne et al., 2016 for a different perspective). Jones, Hills, and Todd (2012) noted that the responses in a free association task represent an *outcome* variable, that is dependent on retrieval processes operating on the cue word's underlying semantic representation. Within this view, performance in the free association task is a dependent variable in and of itself, and free association represents an instance of attentional retrieval from semantic memory, where a specific response has to be selected amongst different activated competitors.

Although conceptualizing semantic memory as a "network" has been primarily explored via free association norms, it is important to mention here that distributional models can also be conceived as semantic networks, such that the angle between word vectors (or cosine similarities) within a vector space could serve as an index of “strength” and be used to create edges (e.g., see Steyvers & Tenenbaum, 2005 for such an approach). In this way, distributional models can be considered “structural models” of semantic memory that provide quantitative estimates for how concepts may be structured and organized within a vector-based semantic space. Therefore, a critical question is whether “structural” models of semantic memory (i.e., DSMs) can explain how free associations are generated. Unfortunately, adequate comparisons of state-of-the-art semantic distributional models in the extent to which they account for free association data itself are limited or lacking. For example, Griffiths, Steyvers, and Tenenbaum (2007) showed that standard DSMs (e.g., Latent Semantic Analysis; LSA; Landaeur & Dumais, 1997) cannot explain asymmetry, violations to the triangle inequality, and the neighborhood structure of free association responses, due to their inherently geometric nature. Specifically, Griffiths et al. showed that cosine similarities between words derived via LSA were inherently symmetric (i.e., *baby-stork* had the same cosine similarity as *stork-baby*), whereas free association norms showed stark asymmetries in producing responses to different cues (i.e., *baby* vs. *stork*), and also did not follow other geometric axioms that LSA (and other geometric DSMs by extension) are bound to follow. Nematzadeh, Meylan, and Griffiths (2017) extended this work to other modern DSMs (word2vec and GloVe; discussed in detail in a later section) to show that these newer models also suffer from the same drawbacks as previous geometric distributional models and demonstrated how a topic model may better account for free association patterns (also see Gruenenfelder, Recchia, Rubin, & Jones, 2016; Jones, Gruenenfelder, & Recchia, 2018). Thus, the research on predicting free associations using DSMs has resulted in mixed findings overall.

It is also noteworthy that most of the work on predicting free associations has been based on the USF norms, which were published in 2004 and contain the first response that comes to mind for 5,019 cues collected from over 6,000 participants. As noted earlier, there now exist more recent and larger databases of free association, which not only measure the first response but also tap into weaker associations using a *continued* free association task, in which participants are asked to produce a certain number of responses that come to mind for a given cue. The largest and most recent such database in English is the Small World of Words database (De Deyne et al., 2019), which contains primary, secondary, and tertiary responses and latencies from over 88,000 participants to over 12,000 cues. A vast body of work has shown that SWOW norms effectively capture similarity judgments (De Deyne et al., 2019), affective and feature-based information (De Deyne et al., 2021), and often outperform distributional models in capturing human behavior (De Deyne, Perfors, & Navarro, 2016). On the other hand, relatively little work has examined performance in the SWOW task from a *predictive* lens; there is a lack of research on how these associations are generated and the factors that influence free association responses and latencies in the SWOW task.

Recently, Thawani, Srivastava, and Singh (2019) used the SWOW norms to construct SWOW-8500, an *evaluation* dataset to compare several DSMs, where they attempted to predict all possible responses to a given cue in a restricted SWOW database using the top *k* neighbors from the DSMs (based on cosine similarity). Specifically, using cosine similarity indices, the top *k* words closest to a given cue were identified within a particular DSM, where *k* corresponded to the total number of unique responses produced by participants to a given cue in the SWOW database. These *k* "predictions" from the DSMs were then compared to the actual SWOW responses and scored for correct and incorrect guesses to compute prediction scores for different

DSMs. Thawani et al.'s work suggested that distributional information could indeed be used to capture SWOW performance (with an average accuracy of about 25%) and showed reliable differences between different DSMs in the extent to which they predicted SWOW responses. In more recent work, Richie, Aka, and Bhatia (in prep.) have applied a neural network-based approach to train a model that learns free association patterns from the SWOW database, by using an error-driven learning approach where vector representations of cues are trained to predict response vectors over several iterations. Although these studies are promising, they are limited in the extent to which they provide a computational account of *how* a particular response is selected for a given cue in the SWOW task. In addition, no studies have attempted to model response latencies, which is a critical aspect of the SWOW norms (that distinguishes this database from the USF norms) and has the potential to uncover important temporal signatures of free association.

1.3 Cloze Task

Similar to the free association task, the sentence completion task (hereafter referred to as the Cloze task; Taylor, 1953) is a widely used laboratory tasks in psycholinguistics and represents another instance of explicit search within semantic memory. In the typical version of the task, a fragment of a sentence (e.g., "The amazing astronaut orbited the") is presented to a group of participants. The participants in the Cloze task are asked to write the word that seems most likely as the next word of the sentence (e.g., *moon, planet*, etc.), and these responses are then normed to produce probabilities for a given response. These probabilities derived from the Cloze task (referred to as Cloze probabilities) are then used to study online lexical/comprehension in several different semantic tasks (Rayner, Ashby, Pollatsek, & Reichle, 2004; Rayner & Well, 1996; Sheridan & Reingold, 2012).

However, similar to free association norms, although there has been extensive work on using Cloze probabilities in other tasks or models, the mechanisms underlying the production of the final word in the sentence fragment itself have not been thoroughly investigated. Of course, producing the final word in the Cloze task clearly involves attending to linguistic and semantic content in the sentence fragment. Therefore, retrieving syntactic/semantic information from underlying representations is presumably critical to this task. Smith and Levy (2011) analyzed the extent to which Cloze probabilities mirror probabilistic estimates of sentence completions from natural language corpora and found medium correlations (ranging from 0.52 to 0.59 across different text corpora), suggesting that the Cloze task may also involve other biases (e.g., familiarity) and processes that are also reflected in reading comprehension times. Of course, this may again reflect the distinction between similarity-driven tasks that distributional models (based on text corpora) are generally good at, compared to production-based tasks that require attending to specific information within the context to select the best candidate.

More recently, Staub, Grant, Astheimer, and Cohen (2015) investigated the mechanisms underlying the Cloze task by examining response latencies to produce the final word. They found that higher Cloze probability responses (based on previously normed data) were produced faster, and more constraining contexts (as defined by the number of total responses to a given fragment) led to faster responses. They interpreted their findings in terms of a race model, in which different Cloze completions raced towards threshold (using simulations), similar to other evidence accumulation models such as the drift-diffusion model (Ratcliff & McKoon, 2008). Clearly, this work is important in attempting to uncover specific mechanisms that may underlie the behavior observed in the Cloze task, although it remains unclear how structural models of semantic memory (i.e., DSMs) account for performance in this task. This is particularly important because recent attention-based DSMs (e.g., BERT, see Chapter 4) are specifically developed from a sentential context perspective. Furthermore, there is limited computational work exploring how different representational models of semantic memory could explain the dynamics of how responses are selected within the Cloze task. Overall, there is a need to explore process-level and algorithmic accounts of production-based semantic tasks such as free association and the Cloze task.

1.4 Algorithmic and Process Models for Semantic Tasks

It is important to distinguish between *process-based* models and what will be called *algorithmic* models. *Process*-based models make explicit mathematical assumptions regarding the flow of information until some decision criterion is reached to drive a response. This dissertation will examine the predictive performance of one such explicit process-based model, the Rotaru, Vigliocco, and Frank (2018) model on various tasks. *Algorithmic* models are much more common in tests of the predictive power of different semantic representational models. As in the original mapping of associative strength to spreading activation in network models (see Collins & Loftus, 1975; Steyvers & Tenenbaum, 2005), these models use variables such as associative strength as a metaphorical index of the amount of activation from one semantic representation to another. Of course, it is quite possible that other metaphors for activation could be used such as simple cosine similarity values used in a Luce-type decision rule (Jones, Gruenenfelder, & Rechhia, 2018) or similarity as an index of featural overlap (Plaut & Booth, 2000). Unless otherwise noted, the present work will use the spreading activation metaphor to estimate the amount of activation between any two concepts within semantic space. As discussed earlier, distributional models allow one to measure the proximity between vector-based semantic representations via cosine similarities, and can therefore be used to develop algorithmic accounts for how responses may be selected within a given semantic task. As we shall see, with the exception of the Rotaru et al model, the models explored here are more algorithmic in nature; this allows one to compare the predictive power of various distributional models in predicting performance.

One way of accounting for performance in production-based semantic tasks is to combine distributional models that explicitly model the learning process to construct semantic representations with explicit algorithmic or process-level assumptions. Although work in this domain is limited, there is some research to suggest that applying well-established algorithms or process-based models can indeed lead to gains in explanatory power for semantic models. As an example of an algorithmic perspective, Jones, Gruenenfelder, and Rechhia (2018) showed how applying the Luce's (1959) choice rule to representations derived from a DSM based on counting co-occurrences of words in a large text corpus and integrating this information with word order information, i.e., BEAGLE (Jones & Mewhort, 2007) can indeed account for performance in the USF norms. Luce's (1959) choice rule estimates the conditional probability of selecting an outcome (e.g., *mango*) given a particular stimulus (e.g., *apple*), by weighting all possible alternatives based on a similarity metric, (e.g., cosine similarity between word vectors derived from BEAGLE) between the cue and the alternatives. Jones et al. showed how this rule (or algorithm) of selecting a response from a set of candidates within a high-dimensional BEAGLE space successfully accounted for patterns of free association responses (violations of symmetry, triangle inequality, and neighborhood structure) within the USF norms that are otherwise problematic for spatial DSMs. In addition to this work applying an algorithmic model based on the Luce-choice rule to predict free associations, some recent work has also attempted to explicitly model temporal dynamics in semantic tasks through a spreading activation framework.

In contrast to more algorithmic models, Rotaru, Vigliocco, and Frank (2018) provide an example of a recent process-based model applied to familiarity-driven tasks. They showed that combining semantic representations derived from three different DSMs, namely, Latent Semantic Analysis (LSA; Landauer & Dumais, 1997), GloVe (Pennington et al., 2014), and word2vec (Mikolov et al., 2013) with a dynamic spreading activation model significantly improved the predictive power of the models on certain semantic tasks. Specifically, Rotaru et al. used a model based on cosine similarity indices derived independently from different DSMs, and modeled the spread of activation within their model as a discrete-time Markov Chain, where activations increased and decreased based on the strength of association between the words before reaching a stable state of equilibrium. To examine how this dynamic activation model predicted accuracy and response times (RTs) for lexical decisions, semantic decisions, concreteness, and imageability ratings, Rotaru et al. evaluated the number of neighbors activated beyond a particular threshold for any particular word and found that these semantic neighbors indeed predicted RTs and accuracy at different timepoints. For relatedness judgments, they found that the strength of association between words in the dynamic model at different time points predicted similarity and relatedness ratings, above and beyond cosine similarities from DSMs. Importantly, this study simultaneously examined the structure of semantic representations derived from DSMs as well as dynamic processes by which these representations are retrieved and brought online during cognitive tasks. Of course, as discussed earlier, it remains unclear whether the processes involved in *familiarity*-based tasks such as lexical/semantic decision or relatedness judgments overlap with those involved in *attentional* tasks such as free association and the Cloze task. In particular, it is possible that familiarity-based tasks may be driven by overall activation within a network (as modeled in Rotaru et al.), whereas more attention-based tasks such as free association and the Cloze task may require selecting an item among several activated competitors, which may instead involve more complex operations, more akin to tasks such as lexical retrieval in speech production.

The primary goal of this dissertation is to simultaneously evaluate different structural models of semantic memory (based on distributional principles) in conjunction with different algorithmic models within two familiarity-driven tasks (relatedness and similarity judgments) and two production-based semantic tasks (free association and the Cloze task), to gain insight into the mechanisms underlying familiarity-based and attention-based retrieval from semantic memory. In addition, investigating response latencies in production-based tasks offers the unique opportunity to understand the temporal dynamics of the search process. Therefore, a related goal of this dissertation is to assess the extent to which different structural models of semantic memory, combined with process-based/algorithmic assumptions account for both response probabilities and response latencies in free association and Cloze tasks, both of which represent different ways of conceptualizing retrieval of a single candidate response from semantic memory. Finally, the present work will also provide a comparison to the Rotaru et al. model as an example of a process-based model, by comparing its performance to the algorithmic models in accounting for relatedness/similarity judgments, free association, and Cloze task performance.

To better understand the differences across the DSMs that will be tested in the empirical sections, the following section provides a relatively brief description of the models (hereafter referred to as structural models). Importantly, the process by which the semantic representation is developed within these structural models turns out to be particularly important for the type of tasks that the models can accommodate.

1.4.1 Structural Models of Semantic Memory

As noted earlier, word-level distributional models, i.e., models that project semantic representations (embeddings) of words onto a high-dimensional vector space, have gained immense popularity in the last few years due to their impressive performance on a variety of semantic tasks such as relatedness/similarity judgments and analogy tasks. A popular word embedding model, word2vec (Mikolov et al., 2013), is a three-layer neural network (NN) model trained to predict a target word in a sentence, given four context words before and after the intended word (continuous bag-of-words version) or vice versa (skip-gram version), using a classifier. By training on millions of context windows in a large text corpus, word2vec tends to develop very rich semantic representations. These have proven to be useful inputs for several downstream natural language processing (Baroni et al., 2014; Collobert & Weston, 2008) and semantic tasks (Mandera et al., 2017), making word embedding models extremely popular in industry and psycholinguistics[2].

Another popular embedding model, Global Vectors (GloVe) was introduced by Pennington, Socher, and Manning (2014). Although GloVe is also an embedding model, in that its semantic representations do project onto a high-dimensional vector space, it is not modeled as a neural network, like word2vec. Instead, GloVe begins with a word-by-word co-occurrence matrix and attempts to estimate the ratio of co-occurrence probabilities between words using a regression

[2] The original Mikolov et al. paper has been cited over 20,000 times as of 2021, as per Google Scholar

model. The primary objective of the GloVe model is to minimize the weighted least-squares error function that emerges from this regression model. The final representations or embeddings that emerge from the GloVe model are particularly sensitive to higher-order semantic relationships. The GloVe model has been shown to perform remarkably well across different semantic tasks; it was originally shown to outperform wordv2vec in analogy tasks and word similarity judgments (Pennington et al., 2014), although more recent work suggests that the performance of the models may depend on the task used to evaluate them (Baroni et al., 2014). Importantly, word2vec and GloVe represent somewhat different proposals for how the meaning of a word may be learned and represented in memory. Specifically, whereas word2vec posits a prediction-based learning mechanism for acquiring word meanings, GloVe focuses on co-occurrence *ratios* within the text corpus, which may emphasize different types of semantic relationships between words (see Discussion for examples). Although word2vec and GloVe have been extensively applied to familiarity-driven tasks (e.g., similarity judgments, verbal analogies, etc.)[3], the extent to which the semantic representations derived from these models explain performance in production-based tasks remains understudied. Therefore, this dissertation evaluates how these structural models (word2vec and GloVe) can be applied to relatedness/similarity judgments (Chapter 2), the free association task (Chapter 3), and the Cloze task (Chapter 4) when combined with the appropriate process-level and algorithmic models.

1.4.2 Modeling Relatedness/Similarity Judgments and Continued Free Associations

The spreading activation mechanism proposes that when a concept or word is activated in memory, its neighbors are also partially activated, which in turn activates other neighboring

[3] Similar to word2vec, the original GloVe paper (Pennington et al., 2014) has been cited over 19,000 times as of March 2021 based on Google Scholar

words in the memory network (Collins & Loftus, 1975). As discussed, this mechanism has been widely applied to semantic tasks (e.g., Neely, 1977) and is considered a central mechanism in computational models of memory (Anderson, 1996). However, computational accounts of how spreading activation may actually be implemented within DSMs have not been thoroughly explored. This section provides a brief overview of the specific process-level and algorithmic models that will be evaluated in this dissertation for the relatedness/similarity judgments and the free association task. Importantly, all models are applied in conjunction with the structural models described above (word2vec and GloVe) to the MEN/SimLex-999 dataset of relatedness/similarity judgments (Bruni, Tran, & Baroni, 2014; Hill, Reichart, & Korhonen, 2014) and the SWOW database of continued free association responses[4].

Given that the SWOW database provides information regarding both first (primary) and second (secondary) responses within an individual, this database allows one to explore distinct models on both the first response and second response. The intriguing question regarding the second response is whether there is any impact of the previously generated first response, even though participants are explicitly instructed to only produce responses to the cue. Therefore, the following section focuses on models that will be applied to relatedness/similarity judgments and *primary* responses in the SWOW database, whereas Chapter 3 describes additional models that will be specifically applied to *secondary* responses in the SWOW dataset.

Rotaru et al. model. As discussed earlier, Rotaru et al. (2018) implemented a *process*-based computational model for lexical and semantic decision tasks, as well as relatedness/similarity judgments. In their model, a pretrained structural model (derived from DSMs such as word2vec and GloVe) was used to obtain vector representations of words, which were in turn used to

[4] I thank Simon de Deyne for providing the response latencies from the SWOW database.

compute cosine similarities between all words and construct a similarity matrix. A series of transformations were then applied to this similarity matrix (details are described in the Methods section of Chapter 2), to create a “dynamic” model that captured activations between different words. Finally, spreading activation was modeled as a discrete-time Markov Chain, which captured the probability of going from one word to another in discrete time steps. The activations between words within this Markov Chain were then used as an indicator of the strength of association between any two given words at a specific time point, which was then applied to predict similarity and relatedness judgments.

Rotaru et al.’s process model provides a general and useful lens through which different parameters that may influence the spread of activation can be explored within the context of semantic retrieval tasks such as relatedness/similarity judgments, free association, and the Cloze task. Therefore, this dissertation applies the Rotaru et al. model to two datasets of relatedness and similarity judgments, as well as the primary responses produced in the SWOW task, to evaluate the extent to which a process-based model may account for relatedness/similarity judgments and free association performance. However, as noted, it is possible that the selection of a single candidate in the free association task may demand different processing assumptions. The attentional demands of retrieving a single candidate may indeed require assumptions that account for the cue’s activation, as well as the activation of competitors within a given task context. The models described below explore the viability of alternative *algorithmic* models in accounting for performance in the relatedness/similarity judgments and free association task.

ELP Baseline Model. As a first step in accounting for variance in different semantic tasks, it is important to consider the influence of simple item-level information contained within the cues and the responses, such as word length, frequency, etc. Therefore, the present work examined the contribution of item-level variables, derived via the English Lexicon Project (ELP; Balota et al., 2007) on relatedness/similarity judgments as well as free association performance. Importantly, all other models were incrementally evaluated against the ELP model, to test whether they explained additional variance over and above baseline item-level characteristics.

Similarity Model. Another account for relatedness/similarity judgments and free association performance relies on the spreading activation metaphor being mapped onto *distance* within a distributional model. Specifically, when a cue is activated, activation spreads to its neighbors based on the *semantic similarity* between the cue and all possible words in the network. In such a model, the structural model space would determine which words are most activated for a given cue, and the word with the greatest similarity to the cue would be selected as the response. It is important to note here that Rotaru et al. showed that their process-based model outperformed the similarity model in predicting relatedness/similarity judgments, although they did not account for the influence of item-level variables in their work. Furthermore, as described earlier, cosine similarities between words within distributional models have been shown to be limited in the extent to which they capture free associations in the USF norms (due to their inherently symmetric nature; Griffiths et al., 2007), although these patterns have not been explored within the more comprehensive SWOW norms. Note that the similarity model is very similar to the top-*k* metric that was used by Thawani et al. (2019) to predict SWOW responses, the difference being that instead of explicitly obtaining top *k* responses from the DSMs (as in Thawani et al.), the similarity model simply measures the cosine similarity for *all* responses produced to a given cue in the SWOW database, and uses these similarities in a predictive regression model in addition to the ELP variables to account for SWOW task performance. Therefore, the similarity model provides a second baseline to compare different algorithmic and process-based models in accounting for relatedness/similarity judgments as well as free associations.

Luce-choice Model. As discussed earlier, Jones et al. showed that applying a Luce-choice decision rule upon similarity estimates derived from a structural model can address the symmetry limitations of the similarity model. Therefore, a Luce-choice model, which estimates the similarity between a cue and a response, conditionalized based on the similarity of the cue to its other neighbors, would predict that when a cue is activated, the activation of given word, *relative* to other activated words, is used to select a response. Given that the influence of a Luce-choice model has not been explored within the SWOW norms, this model is also considered as a viable algorithmic account of free associations. In addition, the present work also explores the contribution of the Luce-choice model in accounting for relatedness/similarity judgments.

Similarity-Frequency Models. Another algorithmic account of free associations may be that when a given cue is activated, activation spreads to its neighbors in proportion to their frequency as well as the semantic similarity to the cue. This type of model assumes that activation spread within a network is driven by not simply the underlying structural space, but also by the frequency of the response and the cue in the language. Moreover, the combined effect of frequency and similarity may be *additive* (i.e., frequency and similarity independently influence response activations) or *multiplicative* (i.e., frequency and similarity interactively influence response activations). Therefore, the present work explores both additive and multiplicative frequency-similarity models as possible accounts for how responses are selected in a free association task, and also applies this model to account for relatedness/similarity judgments.

Multiplicative Similarity-Frequency Delta Models. In addition to the interaction between frequency and similarity, it is possible that individuals are also sensitive to other competitors within the semantic space and the *difference* in activations between a given response and its strongest competitor (i.e., delta) determines the likelihood of selecting a particular response. Therefore, the present work explores whether an additional process that identifies and incorporates the competitor activations, as well as average activations of neighbors of the cue above a certain threshold, provides a better account of performance in the relatedness/similarity judgments as well as the free association task.

1.5 Overview

The first study (Chapter 2) evaluates the extent to which different algorithmic models (discussed above) and one process model (Rotaru et al., 2018), when combined with different structural DSMs (word2vec and GloVe), account for human-generated relatedness and similarity judgments in two publicly available datasets widely used in machine learning and natural language processing (SimLex-99 and MEN). The second study (Chapter 3) evaluates the extent to which these models account for primary and secondary responses and latencies in free association. Again, the question is whether these different DSMs along with distinct different process or algorithmic assumptions can account for responses and response latencies in the Small World of Words database, a *production*-based language task. Further, in order to compare the reliability of the models in explaining free association performance across different datasets, the same models are also applied to an overlapping large subset of the USF dataset of free associations. Finally, Chapter 4 evaluates how structural and process-level/algorithmic assumptions in different DSMs account for Cloze task performance. The data for this study was obtained from Staub et al. (2015; Experiment 2)[5]. It is important to reiterate that structural

[5] I thank Adrian Staub for making this data available

models and process-level/algorithmic accounts are task-dependent, and therefore may not directly apply to other tasks. Therefore, the fourth chapter also describes a very recently developed structural model specifically designed to account for sentence-level performance (BERT, see Devlin, Chan, Lee, & Toutanova, 2019[6]) as well as different algorithmic models that may explain the process of selecting a single response in the Cloze task.

In sum, this dissertation compares different structural models and algorithmic models in the extent to which they account for performance in relatedness and similarity judgments (Chapter 2), free association (Chapter 3) and the Cloze task (Chapter 4). Taken together, these studies provide a quantitative framework to model search and retrieval processes underlying familiarity-based and production-based semantic retrieval tasks, and more generally provide novel insights into the interactions between structure and process in semantic retrieval tasks.

[6] Although the original BERT paper (Devlin et al., 2019) was only published three years ago, it has already been cited over 16,000 times as of March 2021, based on Google Scholar citations

Chapter 2: Modeling Relatedness and Similarity Judgments

There has been considerable emphasis on using computational models of semantic memory to capture variance in tasks that would appear to demand more familiarity-based (non-analytic) processes such as similarity judgments, meaning relatedness, and lexical decision performance. In this chapter, two datasets of human-generated similarity and relatedness judgments that involve more familiarity-based decisions will be evaluated on the common set of process and algorithmic models discussed earlier. It is important to note here that the Rotaru model was designed to account for variance within these specific tasks and is therefore expected to perform well on these datasets. However, it remains unknown whether alternative algorithmic models would be able to account for performance in tasks that are likely driven by overall activation within a semantic space. Therefore, the goal of this chapter is to compare the performance of the process-based Rotaru et al. model with different algorithmic models in accounting for relatedness and similarity judgments.

2.1 Methods

2.1.1 Datasets and Exclusion Criteria

Two datasets were targeted for these analyses, MEN (Bruni, Tran, & Baroni, 2014), and SimLex-999 (Hill, Reichart & Korhonen, 2015). The MEN dataset contains *relatedness* judgment scores for 3000 English word-pairs, where participants were shown two word-pairs and asked to select the more related word-pair on each trial. Each pair was rated against 50 comparison pairs (randomly selected from the same set of 3000 items) by 50 different participants, thus producing an absolute relatedness score on a 50-point scale. For example, *sun-sunlight* produced a perfect relatedness score of 50 (i.e., it was always selected as the more related pair against all 50 comparison pairs), whereas *bakery-zebra* produced a relatedness score of 0 (i.e., it was never selected as the more related pair against all 50 comparison pairs).

SimLex-999 contains *similarity* scores for 999 word-pairs, obtained by asking 500 participants to rate how similar two given words were on a 7-point scale, and then linearly mapping these scores to an 11-point scale (0 to 10). Each word pair was rated by approximately 50 participants. Importantly, participants in SimLex-999 were instructed to specifically focus on similarity and not relatedness, by showing examples of words that may be related (e.g., *car-wheels*) but not similar (e.g., *glasses-spectacles*). Therefore, participants typically rated words that were highly related (e.g., *word-dictionary*, *woman-man*, *dog-cat*, etc.) lower than words that were synonymous (e.g., *vanish-disappear*, *area-region*, *quick-rapid,* etc.) within the SimLex-999 dataset.

For all models evaluated below, an 11,906-word semantic vector space was assumed, which was based on the 11,906 unique one-word cues in the SWOW database, to ensure maximum comparability across different chapters. After converting plural forms of words to singular forms in both datasets (e.g., *daffodils* to *daffodil*, etc.), 17 words in the MEN dataset and 4 words in the SimLex-999 dataset were not within the 11,906 word-space and therefore the MEN dataset was reduced to 2885 word-pairs and the SimLex-999 dataset was reduced to 995 word-pairs[1]. The present study then evaluated the extent to which the two structural models

[1] Overall patterns do not change upon inclusion of these additional words; therefore, results are reported from the restricted dataset to ensure comparability across the similarity/relatedness judgments task and free association task

(word2vec and GloVe) combined with the process model proposed by Rotaru et al. and the various algorithmic models (described in Chapter 1) accounted for relatedness and similarity scores in the MEN and SimLex-999 datasets.

2.1.2 Structural Models

For all analyses, a pretrained word2vec model (skip-gram version) was used which was available from Yamada et al. (2018). The model was trained on a large English Wikipedia corpora (3 billion tokens; extracted in 2018), and produced 300-dimensional word vector representations. A comparable pretrained GloVe model, also trained on a Wikipedia corpus (extracted in 2014; trained on an additional Gigaword 5 corpus; 6 billion tokens), available via Patel, Sands, Callison-Burch, and Apidianaki (2018) was also used to derive 300-dimensional word vectors[2]. These models were then used to obtain vector representations for all words in the MEN/SimLex-999 datasets.

2.1.3 Algorithmic and Process Models

A series of models (introduced in Chapter 1) were implemented using the DSMs described above, to account for similarity and relatedness judgments. The following section describes the mathematical formulations of the different models applied to model relatedness and similarity judgments.

Baseline ELP Model. First, a baseline model was implemented that simply captured similarity of the two word-pairs' item-level characteristics. Specifically, it is possible that words that share item-level characteristics such as frequency, concreteness, valence, etc. are judged to

[2] Pretrained models were used to reduce computational overhead, although future work will control for corpora-level differences. The General Discussion also addresses some of these model differences in detail.

be more similar and/or related. To examine this possibility, item-level information was extracted from the English Lexicon Project (Balota et al., 2007) and included length, frequency, concreteness, and emotional valence for each word pair in the MEN and SimLex-999 datasets. Next, an ELP model was implemented, which contained interaction terms between cue-response item-level characteristics, to evaluate the influence of item-level information on similarity/relatedness judgments. This model provides a baseline for the amount of variance that is simply accounted for by basic item-level characteristics, and therefore does not rely on the semantic information contained within the distributional semantic models. The present study also examined the extent to which ELP-based variables correlated with distributional similarity, to further confirm that these variables likely reflect different sources of information.

Similarity Model. Second, a baseline Similarity (S) model was obtained, based on the cosine similarity between the words vectors. Specifically, using the structural DSMs (word2vec and GloVe), 300-dimensional vector representations for a dataset of 11,906 words were obtained. This matrix, denoted by *vecs*, was of size 11,906 x 300, where each row corresponded to the 300-dimensional vector associated with a particular word. Next, a similarity matrix S, of size 11,906 x 11,906 was computed, from the word vectors, using vector cosine as a measure of similarity between vectors, such that

$$S = (\mathit{vecs}/||\mathit{vecs}||) * (\mathit{vecs}/||\mathit{vecs}||)^{T},$$

where ^{T} denoted the matrix transpose, $||\cdot||$ denoted the Euclidian norm (computed for each row), and / denoted element-wise division. Therefore, S(word1, word2) was computed as the cosine similarity between the two words for each word pair in MEN/SimLex-999 within the specific structural DSM. Importantly, the present work tested whether the Similarity model explained any variance over and above the ELP model described earlier by incrementally adding the cosine similarity estimate to the ELP model.

Rotaru et al. Model. Third, the Rotaru et al. process model was implemented to account for relatedness and similarity judgment scores. Following the procedures described in Rotaru et al., all the negative values in the similarity matrix (S) were set to zero, to construct SM, such that SM(i,j) = S(i,j), if S(i,j) > 0, and SM(i,j) = 0, otherwise. Rotaru et al. assumed that the activation that propagated from the source word w_i to the target word w_j was proportional to both the current activation level of w_i, and the value of SM(i,j), i.e., the strength of the relationship between w_i and w_j. Rotaru et al. further assumed that the total activation within the network remained constant, and therefore, the activation of every word (i.e., row) was merely a sum of the activations of all its neighbors. Therefore, the diagonal elements of the matrix SM were set to zero, and then the rows of SM were normalized, such that each row summed to one, to construct SM_{norm}. Therefore,

$$SM_{norm}\ (i,j) = 0, \text{ if } i = j, \text{ and}$$

$$SM_{norm}\ (i,j) = SM(i,j)\ /\ \Sigma_k\{SM(i,k) \mid 1 \le k \le N \text{ and } k \ne i\}, \text{ otherwise.}$$

Within SM_{norm}, each row represented the conditional probability distribution over all the neighbors of the word associated with that row. To account for the fact that each word may also retain some of its own activation at each time step, Rotaru et al. employed a weighted sum of SM_{norm} and the identity matrix of size N, $I_{N,}$ as the dynamic model, DM = $(2 * SM_{norm} + I_N)/3$. The spreading of activation was modeled as a discrete-time Markov Chain (MC), such that DM represented the probability matrix for the MC, impacted both by the activation of the word itself (I_N) and the strength of activation of its neighbors (SM_{norm}). DM_k denoted the state of MC at step k. This state was computed by raising DM to the power of k, meaning that $DM_k = (DM)^k$. Thus, for any row i and column j, the value DM_k (i,j) represented the probability that the Markov chain was in state j, at time step k, given that it started in state i. This probability denoted the amount of activation associated with word w_j, at time k, following the initial presentation of word w_i. The values DM_k (i,j) and DM_k (j,i) were used to estimate the strength of association between w_i and w_j, and between w_j and w_i, respectively. Therefore, the activation between any two words (i.e., between the word-pair in MEN/SimLex-999) was computed at each time point, resulting in five different process-models, DM_1 to DM_5, each indicating the strength of association between the two words at time steps k = 1 to 5. Next, as in Rotaru et al., these estimates (DM_1 to DM_5) were incrementally added to the ELP + Similarity model (S) within linear regression models to evaluate whether the inclusion of these activations improved the explained variance. Note that although the activation at each time step was incrementally added to the regression model, the final results are reported for the best-fitting model (estimated based on model likelihoods for models k=1 to 5; referred to as the Rotaru et al. model throughout) for brevity.

Luce-Choice Algorithmic Model. The algorithmic Luce-choice model investigated the possibility that relatedness/similarity decisions are made by examining the *relative* similarity between words, based on the Luce-choice decision rule. Although simpler models of the Luce-choice rule exist (e.g., Jones et al., 2018), the frequency-biased version of the Luce-choice model was implemented for these analyses for maximum comparability with the subsequent models[3]. Specifically, in the Luce-choice (Luce) model,

$$\text{Luce}(\text{word}_2 \mid \text{word}_1) = F(\text{word}_2)*S(\text{word}_2, \text{word}_1)/\ \Sigma_k\{F(k)*S(k, \text{word}_1),$$

[3] For Luce and SF-based models, all negative cosine similarities were set to a small positive value (.0001) to ensure comparability with the Rotaru et al. model and ensure that multiplicative functions did not yield zero products.

where $word_1$ and $word_2$ denoted the first and second words presented to participants in the MEN/SimLex-999 datasets, *k* ranged from 1 to *tau* and denoted the topmost *tau* neighbors of the first word, and F ($word_2$) denoted the spoken word frequency of the second word. Log frequency estimates were obtained using the SUBTLEX-WF database available from the ELP[4] and then activations were computed for each word within the Luce-choice model. Although initial analyses explored the parameter space for *tau* (by counting all words with activations above a certain number of standard deviations as the neighbors of the first word), setting *tau* = 11,906 (i.e., the complete similarity space) produced the best results, and therefore the final results are reported only for *tau* =11,906. Of course, the idea of "neighbors" is relative here, and given the continuum of similarities within vector-space models, all words are technically neighbors of each other, albeit by varying degrees (as indicated by cosine similarities). For example, the word *dog* is highly similar to *puppy,* slightly less similar to *cat*, and least similar to *apple*. Within this view, the *relative* similarity of a given word ($word_2$) to another word ($word_1$), conditionalized by *all* possible similarities in the semantic space is assumed to influence relatedness and similarity judgments. Given that the second word's frequency is already incorporated into the Luce-choice model formulation, this term was excluded from the ELP model for these analyses to avoid double-dipping (the first word's frequency, and all length, concreteness, and valence-based terms were retained).

Similarity-Frequency Additive/Multiplicative Models. The Similarity-Frequency (SF) models explored the contribution of frequency in determining the similarity and relatedness scores. Specifically, it is possible that relatedness or similarity judgments are influenced by the

[4] Frequency estimates for 476 words out of 11,906 total words were missing from the ELP -- to ensure the $SF_{multiplicative}$ was symmetric and did not have missing values, log frequency for the missing words was set to 1 (mean log frequency was 2.37 (*SD* = .82) across all words). Patterns did not change if these words were excluded from analyses.

frequency *and* semantic similarity of the different words within the semantic space, and this combined effect could be additive or multiplicative. Therefore, in the SF models,

$$SF_{additive}(word_2 \mid word_1) = S(word_1, word_2) + F(word_2)$$

$$SF_{multiplicative}(word_2 \mid word_1) = S(word_1, word_2)*F(word_2)$$

As in the Luce-choice model, given that frequency of the second word is already incorporated into the SF models, the interaction term for word frequencies from the ELP model was excluded for these analyses, to avoid double-dipping. Note that the $SF_{multiplicative}$ model is quite similar to the Luce-choice model, except that it doesn't take into account the *relative* similarity to other responses.

Multiplicative Delta Model. The multiplicative delta model ($SF_{multiplicative}$-delta) explored whether adding an additional step of accounting for neighboring activations improved the predictive power of the $SF_{multiplicative}$ model. Specifically, it is possible that similarity and relatedness judgments are influenced by the activation of other words within the network, such that the *difference* in activation of the specific word ($word_2$) and the next most active competing word may determine the extent to which two words is considered to be related or similar. To evaluate this possibility, for each word pair, the difference (delta) in activations indexed by the cosine*frequency values between the second word ($word_2$) and the next most active word within the semantic space for the first word ($word_1$) was computed, and these estimates were used to predict relatedness/similarity scores within a regression model. Importantly, it is possible that delta reflects a type of *thresholding* process, such that when a word is sufficiently more activated than a competitor beyond a threshold, it is considered to be related/similar. This may suggest that the contribution of delta asymptotes at some point, or only starts to influence response likelihoods after a certain level of activation. Therefore, the present study also examined whether delta showed a quadratic trend with relatedness/similarity scores (which would be suggestive of a threshold), and included a quadratic term in the regression models if the quadratic trend was significant. If the quadratic trend was not significant, a linear delta term was added to the models.

Multiplicative Delta-Neighbors Model. In addition to the competing activations of a single competitor, it is also possible that the overall activation of neighbors within a semantic space influences similarity/relatedness judgments. To evaluate this possibility, the mean activation values of neighbors of the first word with cosine*frequency values above 3 standard deviations within $SF_{multiplicative}$ were computed and included as an additional predictor within the $SF_{multiplicative}$-delta-neighbors model. Although initial models explored the full spectrum of neighbors (e.g., neighbors above 1, 2, 4, etc. standard deviations), these analyses revealed similar patterns but low overall variance, and have therefore not been reported.

2.2 Results

For all analyses, total explained variance (R^2) computed using the *r.squaredGLMM* function from the MuMIn package in R (Barton, 2020)[5] was used to estimate the predictive power of the different models, and the *Weights* function from the MuMIn package was used to estimate model likelihoods. Specifically, the relative evidence in favor of one model versus another was assessed using normalized model likelihoods obtained by supplying Bayesian Information Criterion (BIC) indices to the *Weights* function. In addition, to assess the variability in the obtained R^2 estimates, bootstrapped confidence intervals were obtained for each fixed-effect R^2 estimate by sampling

[5] The r.squaredGLMM function gives marginal and conditional estimates of R^2 for mixed-effects models, and a single estimate for linear regression models. Both estimates are reported wherever mixed-effects models are used.

with replacement across 1000 simulations using the *boot* function in R. Further, the *anova* function was used to test significance of nested models, and the cocor package (Diedenhofen & Musch, 2015) was used to compute statistical significance of differences between correlations across all analyses.

Table 2.1 displays the total explained variance in relatedness and similarity scores in the MEN and SimLex-999 dataset, respectively. First, it is noteworthy that the ELP base model explained considerable variance across both datasets, indicating that words with similar item-level characteristics were considered higher in similarity and relatedness compared to words with more dissimilar item-level characteristics overall. For example, pairs with low concreteness words such as *weird-normal* were scored as less similar (SimLex-score = 0.72), whereas pairs with high concreteness words such as *horse-mare* were scored as more similar (SimLex-score = 8.33). Further, pairs with high valence words such as *happy-smile* were judged as more related (MEN-score = 40), whereas word pairs with low valence words such as *flood-line* were judged as less related (MEN-score = 26) overall. Importantly, these item-level characteristics showed only moderate correlations with cosine similarities derived from word2vec and GloVe ($-.09 \leq r \leq .20$), indicating that these relationships did not entirely overlap with distributional information. Next, as shown in Table 2.1, different algorithmic models and the Rotaru et al. model were examined in the extent to which they explained variance in relatedness/similarity judgments. It is important to highlight here that all models' estimates were added over and above the predictors from the ELP model, and therefore one could assess whether these models explained significant variance over and above baseline item-level characteristics.

First, as shown, explained variance was significantly higher for MEN than for Simlex-999, consistent with previous work, which may be indicative of differences in the demands of the task as well as the nature of items used in both datasets (De Deyne et al., 2019; Rotaru et al., 2018). Importantly, the Rotaru et al. model explained the most variance across both datasets, although bootstrapped confidence intervals slightly overlapped between the Rotaru et al. model and the simple similarity model in GloVe-based models, suggesting that there was some variability in the explained variance estimates. Furthermore, word2vec and GloVe appeared to explain similar amounts of variance across both datasets, although overall, model likelihoods based on BIC indices indicated that the Rotaru et al. model based on GloVe clearly performed better than word2vec across MEN and SimLex-999[6].

Table 2.1. Explained Variance in MEN and SimLex-999

Structural Model	**Process Model**	**MEN: R^2 [CI] (%)**	**SimLex-999: R^2 [CI] (%)**
	ELP	10.53 [7.62, 12.46]	16.86 [11.92, 19.68]
word2vec	Similarity*	55.51 [52.83, 57.66]	28.05 [22.40, 31.83]
	Rotaru et al. model*	**61.47 [58.59, 68.63]**	**32.94 [27.23, 36.35]**
	Luce	47.88 [44.71, 50.49]	20.96 [15.97, 24.37]
	$\text{SF}_{\text{additive}}$	14.15 [11.01, 16.41]	15.90 [11.23, 18.70]
	$\text{SF}_{\text{multiplicative}}$	45.53 [42.33, 48.12]	20.35 [15.42, 23.71]
	$\text{SF}_{\text{multiplicative}}$-delta	48.13 [44.75, 50.79]	20.98 [15.89, 24.21]
	$\text{SF}_{\text{multiplicative}}$-delta-neighbors	52.32 [49.03, 54.87]	23.86 [18.53, 26.83]
GloVe	Similarity*	62.55 [59.96, 64.70]	32.44 [26.71, 36.42]
	Rotaru et al. model*	**64.88 [62.31, 66.88]**	**34.55 [28.87, 38.25]**
	Luce	49.55 [46.49, 52.09]	23.66 [18.58, 27.26]
	$\text{SF}_{\text{additive}}$	15.48 [12.27, 17.84]	15.86 [11.21, 18.66]
	$\text{SF}_{\text{multiplicative}}$	50.55 [47.48, 53.07]	22.01 [16.99, 25.51]
	$\text{SF}_{\text{multiplicative}}$-delta	52.01 [48.76, 58.60]	22.09 [16.91, 25.31]
	$\text{SF}_{\text{multiplicative}}$-delta-neighbors	56.09 [52.86, 58.56]	26.95 [21.41, 30.19]

Note: * indicates significant ($p < .05$) increase in variance based on log-likelihood tests over the previous model. CI indicates the 95% confidence interval based on bootstrapped R^2 estimates from 1000 samples with replacement.

[6] In Rotaru et al., word2vec outperformed GloVe in both MEN and SimLex-999, although these differences could be attributed due to differences in text corpora given that Rotaru et al. used the British National Corpus, while the present work uses the Wikipedia corpus. Additionally, Rotaru et al. did not account for ELP variables.

Interestingly, the algorithmic models (i.e., Luce and SF-based models) did not adequately explain similarity and relatedness judgments in MEN and Simlex-999 (although Luce and $SF_{multiplicative}$-based models explained more variance than the baseline ELP model across both word2vec and GloVe)[7]. Overall, these results indicate that the Rotaru et al. process model successfully explained relatedness and similarity ratings considerably better than the algorithmic models. Of course, this was expected given that the Rotaru et al. model was specifically developed to account for performance in familiarity-based tasks, and the present findings extend the Rotaru et al.'s process-model to a new corpus and also show that the Rotaru et al. model explains significant variance over and above shared item-level characteristics between the words.

2.3 Discussion

The results from the current study yielded five important observations. First, it is noteworthy that the ELP variables accounted for a substantial amount of variance in these tasks, and hence future studies comparing predictive power for different models should control for simple item-level variables. Another noteworthy observation is that the ELP model accounted for much more variance in SimLex-999 than MEN, but this was opposite to the patterns observed from the algorithmic and process-based models, where explained variance was higher in MEN compared to SimLex-999; this too may reflect the differences in task demands and the extent to which item-level information is accessed across MEN and SimLex-999. Specifically, the task of selecting the more related word-pair between two pairs (as in MEN) likely involves different processes than rating a single word-pair (as in SimLex-99). Indeed, the single-pair judgment is likely to be more sensitive to lexical biases compared to the two-pair judgment, which may

[7] Delta did not show a significant quadratic trend (*p* > .05); therefore, all delta terms reflect linear relationships.

involve more attentional processing. The task-specific demands of MEN vs. SimLex-999 are therefore critical, and previous studies comparing model performance on these datasets (e.g., De Deyne et al., 2019; Rotaru et al., 2018) have not controlled for such lexical biases, which may have influenced the outcomes and interpretations of these studies.

Second, as expected the Rotaru model captured the most variance in both datasets. Third, remarkably, the simple Similarity model performed almost as well as the Rotaru et al. model, which suggest that participants were indeed relying on simple proximity in semantic space to make judgments in these tasks. Fourth, the competition between words and competitors (captured via delta-based models) did not appear to help and in fact decreased variance compared to the simple similarity model. Finally, GloVe explained slightly more variance than word2vec, however, this is confounded a bit by differences in the training corpora across the DSMs.

Overall, the first study explored how a process model based on the spreading activation mechanism, as well as different algorithmic models explained relatedness and similarity judgments. The critical finding from this study was that the spreading activation-based process model proposed by Rotaru et al. successfully predicted similarity and relatedness judgments in the SimLex-999 and MEN datasets, and outperformed other algorithmic models in these tasks. These results suggest that tasks that may be driven by summed activations within a semantic space can indeed be successfully modeled by such a process model.

In addition, the present results indicated that GloVe performed better than word2vec on across both datasets predicting similarity and relatedness, although bootstrapped confidence intervals for the two DSMs overlapped, indicating that there was considerable variability in these estimates. Although previous work in natural language processing has shown that word2vec-type models generally perform better than GloVe on SimLex-999 but not on MEN[8], it is important to note here that the models in the current study were different in a few ways from previous work on MEN/SimLex-999. First, the current analyses estimate total explained variance as well as bootstrapped confidence intervals that provide more information about the variability of these variance estimates. Second, the present analyses also accounted for the influence of item-level characteristics via the ELP model in addition to cosine similarities derived from word2vec and GloVe, in an effort to estimate the total variance explained in the tasks when controlling for lexical characteristics. Given that ELP variables correlated only moderately with cosine similarities, these results indicate that the present GloVe model does perform better than word2vec on MEN/SimLex-999 when accounting for item-level characteristics. An important point to mention here is that the GloVe model used in the present study was trained on an additional corpus (Gigaword 5), which may have contributed to some of these patterns. However, it is also possible that the differences observed in the predictive power of word2vec vs. GloVe reflect the *type* of information that the two models tend to capture. Indeed, word2vec is a neural network trained to predict words that follow other words within a 4-word context window, whereas GloVe attempts to capture meaningful co-occurrence-based relationships between words within the text corpus. Specifically, GloVe estimates co-occurrence *ratios* for different words across the full corpus, whereas word2vec uses the co-occurrence *counts* to predict words within context windows. It is possible that estimating *ratios* of co-occurrence allows GloVe to not only encode which words are *similar* to each other, but also how different words may be *related* to each other. Consider the example provided by Pennington et al. (2014). The words *ice*

[8] See state-of-the-art results on MEN (https://aclweb.org/aclwiki/MEN_Test_Collection_(State_of_the_art)) and SimLex-99 (https://aclweb.org/aclwiki/SimLex-999_(State_of_the_art))

and *steam* are both related to *water* and frequently co-occur with *water*, but this relationship is not useful in differentiating the meaning of the two words. However, *ice* is more related to *solid* than *gas* and *steam* is more related to *gas* than *solid*, and therefore the ratio of co-occurrence between *solid-ice* and *solid-gas* could be informative about the specific properties of *ice* vs. *steam*. Therefore, GloVe predicts these co-occurrence ratios, and assumes that words that have higher co-occurrence ratios (e.g., *solid*) are more *related* to the one word (e.g., *ice*) vs. another (e.g., *steam*). Note that this is process is quite different from word2vec's process of predicting which words may fit a given sentential context, which may rely more on identifying which words are used within similar syntactic positions across different linguistic contexts. Therefore, it is possible that attending to co-occurrence ratios allows GloVe to capture different types of semantic relationships, compared to word2vec, which may be more biased towards similarity-type relations. Indeed, *relatedness* is a somewhat broader construct, and *similarity* is often considered a special case of relatedness (De Deyne et al., 2019). Hence, due to capturing these co-occurrence ratios, GloVe may be simply better at capturing different forms of relatedness *and* similarity, which gives it an advantage over word2vec after having accounted for basic item-level characteristics. Of course, these hypotheses are post-hoc, and future work should perform more focused tests of predictive power using controlled corpora, to fully understand the predictive power of the different models and the underlying processes that govern relatedness and similarity judgments. Indeed, recent work suggests that such judgments of semantic relatedness may be governed by decision-based processes (Kraemer, Wulff, & Gluth, 2021) that can be captured by computational process models such as the leaky accumulator model (Usher and McClelland, 2001). Although MEN and SimLex-999 did not contain response latencies, future work should also examine the extent to which different process-level accounts can accommodate important temporal signatures in similarity and relatedness judgments.

In sum, the present study evaluated different algorithmic and one process-based model in accounting for similarity and relatedness judgments derived from MEN/SimLex-999, and showed that the Rotaru et al. model based on the spreading activation mechanism provided the best account for these data. Although tests of predictive power for different DSMs in such familiarity-based tasks are fairly common, there has been very little work exploring the predictive power of different computational models of semantic memory on production tasks. Therefore, Chapters 3 and 4 focus on two such production-based tasks, free association and the Cloze task, and evaluate the extent to which different models can account for responses and response latencies in these more attention-demanding retrieval tasks.

Chapter 3: Modeling Continued Free Association Responses and Latencies

Associative strengths derived from free association responses have been widely applied to understand different cognitive phenomena, as well as develop computational network models of semantic memory. However, as discussed in Chapter 1, the mechanisms underlying free association continue to remain unclear. This study evaluates how two structural models (word2vec and GloVe), when combined with different algorithmic models and one process model (Rotaru et al., 2018), predict primary and secondary response proportions as well as RTs in a continued free association task (SWOW; De Deyne et al., 2019). Importantly, although prior work has examined free association responses within the USF database (e.g., Griffiths et al., 2007; Jones et al., 2011), no work has investigated RTs to produce the associate response or the secondary responses that are available in the SWOW database. Investigating the variables and processes that influence RTs and secondary responses in the free association task may provide insights into the dynamics of this task and could potentially provide a better index to evaluate different models. Further, the present study also examines how different models account for performance in the USF dataset to provide converging evidence that the patterns observed in the SWOW task do indeed replicate in an older lab-based dataset.

3.1 Methods

3.1.1 Dataset and Exclusion Criteria

Small World of Words Dataset. The dataset of free association responses collected by De Deyne et al. (2019) in the Small World of Words (SWOW) project was the primary dataset for these analyses. The SWOW data come from an online task[1] that involves producing words that come to mind in response to a given cue (see https://smallworldofwords.org/en). Participants were presented a cue word on the screen and instructed to respond with the first three words that came to mind. They were also instructed to respond only to the cue word (and not to previous responses), and could press a “no more responses” button if they could not think of further responses to a given cue. The SWOW dataset[2] contains primary, secondary, and tertiary word associations and response latencies for 12,292 cue words in English, produced by 101, 892 participants.

Due to the online nature of this task, there was considerable variability in how participants approached the task (e.g., participants could produce responses to only one or many words at their own pace, and used different devices to perform the task). Therefore, a series of selection criteria were applied to the raw dataset to obtain a quality dataset to model. First, 3.17% of the SWOW dataset contained cues that were longer than one word (e.g., *thank you, far away, high school,* etc.), which were excluded to ensure that vector representations for these words could be derived from the word-level structural models (word2vec and GloVe). The reduced dataset contained 11,906 unique one-word cues. Second, any responses that were not presented as cues (to ensure that model computations were on symmetric matrices; as is standard in free association research, see Nelson et al., 2004) were excluded, which resulted in 11,750 unique primary responses and 11,748 unique secondary responses. Third, trials on which

[1] The SWOW project was initially based on a pen-and-paper task (De Deyne et al., 2019), but the current dataset is exclusively from the web-based version of the task, in which individuals could participate via computers and mobile devices.

[2] Shared by Simon De Deyne.

negative RTs (likely reflecting server issues or poor connectivity) were reported were excluded (0 trials for primary responses and 53,353 trials for secondary responses), and to further ensure good estimates for RTs, only responses that were produced by at least 4 participants were retained[3]. After implementing these procedures, the final dataset of primary responses (referred to as SWOW-R1) contained 717,736 observations from 98,528 participants, with 11,903 unique cues and 8,737 unique primary responses. For secondary responses, given that the algorithmic models critically depended on the primary response, responses which did not fall within the original 11,906 cues were further excluded. The final dataset of secondary responses (referred to as SWOW-R2) contained 373,964 observations from 89,249 participants, with 11,861 unique cues and 7,771 unique secondary responses. All reported analyses are based on these final datasets of primary and secondary responses (SWOW-R1 and SWOW-R2).

USF Dataset. In addition to examining the predictive power of different structural and algorithmic models in accounting for performance in the SWOW task, the present study also compared the patterns observed in the SWOW task to the gold standard free association norms, collected by Nelson et al. (2004). As discussed earlier, the USF norms were collected in a lab-based setting from over 6,000 participants for 5,019 words with an average of 150 responses per cue. To compare the two datasets in the fairest way possible, a subset of the USF norms was considered, which excluded responses produced by less than 4 participants (as in SWOW-R1) as well as any cues or responses that were not included in the 11,906 words, leading to a total of 4,985 unique words. Next, activation estimates from different structural and algorithmic models were obtained within this reduced dataset (USF-4985), and these estimates were compared to a

[3] Variance explained was very low when responses with less than 3 participants were included. Note that for secondary responses, primary responses produced by less than 4 participants were retained as long as the secondary response itself was produced by at least 4 participants.

smaller subset of SWOW-R1 which contained R1 responses to the same 4,985 cues also present in USF-4985 (hereafter referred to as SWOW-4985).

3.1.2 Structural and Algorithmic/Process-based Models

As in Chapter 2, pretrained word2vec and GloVe models were used to obtain 300-dimensional vector representations for all cues and responses in the SWOW-R1 and SWOW-R2 databases. Next, a series of models were evaluated in the extent to which they explained performance in the free association task. Importantly, given that the SWOW dataset contains both primary and secondary responses, the models implemented differed for the primary and secondary responses and are therefore described separately below.

Primary Response Models.

ELP model. The ELP model examined the extent to which basic item-level characteristics of the cue and response influenced free association responses and latencies. Specifically, it is possible that words that are more concrete or have high emotional valence or frequency tend to also produce responses that are concrete, have high emotional valence, or frequency. Therefore, as in Chapter 2, item-level information for cues and responses was extracted from the ELP database and submitted to a regression model predicting primary responses and response latencies.

Similarity Model. As before, the similarity model examined the extent to which cosine similarity from the DSMs (word2vec and GloVe) accounted for free association performance. Therefore, S(cue, response) was computed as the cosine similarity between the cue's vector and the response vector within the specific structural DSM and submitted to a regression model including the ELP variables, to evaluate the extent to which semantic similarity explained additional variance in free associations over and above item-level information.

Rotaru et al. Model. As in Chapter 2, the Rotaru et al. model was implemented to obtain activations from the cue to the response at discrete time steps, and these activations were added to the ELP + Similarity model to evaluate whether the process-based model explained additional variance in free associations over and above baseline item-level information and cosine similarity between the cue and response.

Luce-Choice Algorithmic Model. The algorithmic Luce-choice model investigated the possibility that responses to cues in the free association task are selected not merely via semantic similarity to the cue, but instead the *relative* similarity of the specific response, compared to other neighbors of the cue, based on the Luce-choice decision rule. Specifically, in the Luce-choice (Luce) model for free associations,

$$\text{Luce(response} \mid \text{cue)} = \text{F(response)*S(cue, response)/ } \Sigma_k\{\text{F(k)*S(cue, k)}\},$$

where *k* was set to 11,906, as in Chapter 2, and F (response) denoted the spoken word frequency of the given response derived via the ELP.

Similarity-Frequency Additive/Multiplicative Models. The similarity-frequency models explored the joint contribution of semantic similarity and frequency on free association performance. Specifically, it is possible that the process of selecting a particular response for a given cue also takes into account baseline activations of the different words within a similarity space, where baseline activations could reflect the frequency of a given word. Therefore, when a cue is activated, it activates other neighbors as a function of both the frequency of the word, as well as its semantic similarity to the cue itself, and this function could be additive or multiplicative.

$$SF_{additive}(\text{response} \mid \text{cue}) = S(\text{cue, response}) + F(\text{response})$$

$$SF_{multiplicative}(\text{response} \mid \text{cue}) = S(\text{cue, response})*F(\text{response})$$

Activations from the SF-based models were submitted to a regression model containing ELP variables[4] to evaluate the extent to which semantic similarity and frequency influence response production.

Multiplicative Delta Model. The multiplicative delta model assumed that once a group of potential responses was identified, the response that was comparatively higher in activation from the next most active response is selected. For example, for a given cue *village*, responses such as *town, city,* etc. may come to mind, and ultimately, the *difference* between the activation of a given response from the next most activated response may determine the selection process. To evaluate this possibility, for every cue-R1 combination, the cosine*frequency value of the next most active competitor within $SF_{multiplicative}$, and the difference between R1's activation and the competitor's activation (*delta*) were computed and these estimates were used in a regression model ($SF_{multiplicative}$-delta) to predict R1 probabilities and RTs. As before, a quadratic term for delta was included within the models if delta showed a significant quadratic pattern with response probabilities and latencies.

Multiplicative Delta-Neighbors Model. In addition to examining the difference in activations between R1 and the strongest competitor, the multiplicative delta-neighbors model examined whether the *mean* level of activations for the strong neighbors of the cue influenced the likelihood of selecting a given response. It is possible that if neighbors of the cue (other than R1 and the strongest competitor) are highly activated on average, this either reduces the overall likelihood of selecting a given R1 due to excessive competition among the different neighbors,

[4] Excluding response frequency, to avoid double-dipping

or facilitates the production of R1 due to converging activation from multiple words. Therefore, as in Chapter 2, the mean neighbor activation for a specific cue was computed and submitted to a regression model in addition to the estimates of similarity-frequency and delta ($SF_{multiplicative}$-delta-neighbors) to predict R1 probabilities and RTs.

Secondary Response Models. In order to model secondary response production in the SWOW norms, there is a need to incorporate additional assumptions that account for the selection of the primary response. Therefore, the present work explores how secondary responses are selected in the SWOW task by providing additional activation to certain words within the semantic space after a specific primary response has been selected, via the *unchained* and *chained* models described below. Importantly, because the $SF_{multiplicative}$-delta-neighbors model provided the best fit to R1 responses (see Results section), which was in turn based on the underlying values in the $SF_{multiplicative}$ matrix, chained and unchained secondary response models were first derived from values within the $SF_{multiplicative}$ model. Competitor and neighbor activations were then subsequently obtained from the $SF_{multiplicative}$ model to ultimately test the influence of competing neighbors in secondary response production via the $SF_{multiplicative}$-delta-neighbors model.

Unchained Multiplicative Model. The *unchained* model assumed that activation from the cue spread to its neighbors excluding R1, assuming that R1 has already been produced or “tagged” as such (Dell, 1986). The idea of tagging or dampening an already produced response is important in theories of speech production, and serves as an indicator that a response has been selected for output and therefore should not be further activated. Therefore, within the unchained model, the “activation” of R2 was simply the value of the specific response in $SF_{multiplicative}$:

$$SF_{multiplicative}\text{-unchained } (R2 \mid cue\text{-}R1) = SF_{multiplicative}(R2 \mid cue)$$

Importantly, the unchained model was further supplemented using the same estimates of delta (indicating the difference in the activation of the response and the strongest competitor) and the mean neighbor activations, via the $\text{SF}_{\text{multiplicative}}$-unchained-delta and $\text{SF}_{\text{multiplicative}}$-unchained-delta-neighbors models to explore whether these additional assumptions about neighbors and competitors improved the fit of the unchained model.

Chained Multiplicative Model. A *chained* model of secondary response production assumed that activation spread from R1 to its neighbors, excluding the cue. Importantly, to simultaneously assess the contribution of the cue spread vs. R1 spread in predicting R2, the amount of additional activation (as indexed by values in $\text{SF}_{\text{multiplicative}}$) provided by the cue (theta) and by R1 (1-theta) to their neighbors above 2 and 3 standard deviations was parametrically varied. Specifically, theta was varied from 0.1 to 0.9 to examine how varying the relative contribution of "activation" from the cue vs. R1 influenced final R2 values in $\text{SF}_{\text{multiplicative}}$-chained-cueR1. Therefore, in the $\text{SF}_{\text{multiplicative}}$-chained-cueR1 model, a particular word could receive additional activation from both the cue and R1, only the cue, only R1, or neither. This formulation resulted in 9 (theta values) x 2 (standard deviation values) = 18 distinct models for each structural DSM (word2vec and GloVe).

Furthermore, in addition to testing the *relative* contribution of the cue vs. R1 in predicting R2 responses as described above, the $\text{SF}_{\text{multiplicative}}$-chained-R1 model examined specific weight of R1's value in $\text{SF}_{\text{multiplicative}}$ (beta) and the total number of neighbors to which this value was added (neighbors of R1 with cosine similarity values above *n* standard deviations, where *n* ranged from 1 to 5) to examine the full extent to which R1 influenced R2 responses. This resulted in 9 (beta) x 5 (*n*) = 45 model configurations for each structural DSM. The estimates from these models were used predict R2 response probabilities and latencies. Furthermore, similar to the unchained model, the SF$_{\text{multiplicative}}$-chained-R1 model was further extended to incorporate delta and neighbor-based activations, yielding the SF$_{\text{multiplicative}}$-chained-R1-delta and SF$_{\text{multiplicative}}$-chained-R1-delta-neighbors models.

3.2 Results

3.2.1 Descriptive Statistics

Overall, there was considerable variability in the responses produced by participants. First, it should be noted that the number of cues that a given participant responded to ranged from 1 to 22 (*M* = 7.28, *SD* = 2.9) in SWOW-R1 and from 1 to 17 in SWOW-R2 (*M* = 4.19, *SD* = 2.15). So, most participants only responded to a few cues in the SWOW datasets. Figure 3.1 shows the distribution of *unique* primary (R1) and secondary (R2) responses across participants to different cues in SWOW-R1 and SWOW-R2.

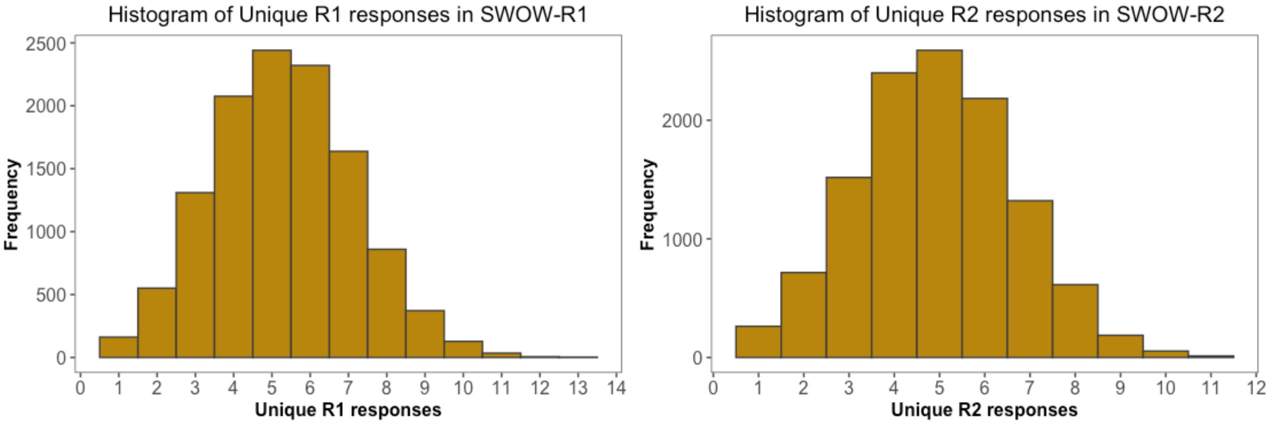


Figure 3.1. Distribution of unique primary (R1) and secondary (R2) responses in the SWOW-R1 and SWOW-R2 databases.

As shown, across all participants in the datasets, the total number of unique R1 responses to the different cues ranged from 1 to 13 (*M* = 5.3, *SD* = 1.86) in SWOW-R1. Thus, some cues produced more varied responses than others. For example, cues such as *affection*, *beagle*, and

Yellowstone produced only 1 unique R1 response (e.g., *love*, *dog*, and *park*, respectively) in SWOW-R1, whereas cues such as *foggy, surgical*, and *waist* produced 13 unique R1 responses (e.g., *band, belt*, *body*, *coat*, etc.). Similarly, the number of unique R2 responses ranged from 1 to 11 in SWOW-R2 (M = 4.92, SD = 1.76). Interestingly, the cues that produced the fewest unique R1 responses did not directly correspond to those that produced fewest unique R2 responses, although there was a moderate correlation between the number of unique R1 and R2 responses for a given cue, r = .27 (p < .001). For example, although *affection* produced only 1 unique R1 response in SWOW-R1, i.e., *love*, it produced 4 different R2 responses in SWOW-R2: *caring*, *hug*, *hugs*, *love*[5]. This is important because it suggests that participants may indeed be using R1 responses and the cue to produce R2 responses, i.e., a type of chaining. Further, the mean number of unique R2 responses for specific cue-R1 pairs was 1.48 (SD = .95), suggesting that several cue-R1 combinations produced singleton R2 responses.

3.2.2 Predicting Primary (R1) Responses in Free Association

To analyze the primary responses, the SWOW-R1 data was first aggregated to obtain probabilities of producing different R1 responses for a given cue. Next, as described earlier, the influence of item-level characteristics that may influence free association performance was investigated via the ELP model. Table 3.1 displays the correlations between the item-level characteristics of the cue and R1. As shown, high cue frequency, concreteness, and valence were positively correlated with high R1 frequency (r = .18 , p < .001), concreteness (r =.61 , p < .001), and valence (r = .56 , p < .001), indicating that cues with high frequency, concreteness, and

[5] Note that as discussed in the Methods section, although SWOW-R1 excluded R1 responses that were produced by fewer than 4 participants, these responses were retained in SWOW-R2 as long as the specific R2 response was produced by at least 4 participants. Therefore, *love* could be the R2 response for a trial on which *like* was the R1 response, even if *like* was not produced by at least 4 participants.

valence produced responses with high frequency, concreteness, and valence, respectively. For example, the cue *you* (a high-frequency word), produced a high-frequency response such as *me* more often (probability = .84) compared to a response with lower frequency such as *person* (probability = .10).

Table 3.1. Correlations between cue-R1 ELP variables in the SWOW-R1 database

	cue-Length	**R1-Length**	**cue-Conc**	**R1-Conc**	**cue-Val**	**R1-Val**	**cue-Freq**	**R1-Freq**
cue-Length	1							
R1-Length	0.2*	1						
cue-Conc	-0.32*	-0.17*	1					
R1-Conc	-0.23*	-0.25*	0.61*	1				
cue-Val	0.01*	0.01*	0.07*	0.06*	1			
R1-Val	-0.03*	-0.02*	0.08*	0.12*	0.56*	1		
cue-Freq	-0.32*	-0.07*	0.08	0.01*	0.17*	0.08*	1	
R1-Freq	-0.02*	-0.43*	-0.05*	0.02*	0.07*	0.21*	0.18*	1

Note: * indicates significant correlation (*p*< .05). Conc indicates concreteness rating, Val indicates emotional valence rating, and Freq indicates SUBTLEX word frequency derived via the ELP

Similarly, the cue *happiness* (a high-valence word) produced a high-valence response such as *joy* more frequently (probability = .44) than *sadness* (a low-valence word; probability = .12). Finally, a concrete cue such as *apple* produced concrete responses such as *pear* and *orange*, whereas a low-concreteness cue such as *spirituality* produced responses with low concreteness such as *religion* more frequently (probability = .78) to responses with high concreteness, such as *church* (probability = .11). There were also some interesting correlations between length and other ELP variables, suggesting that longer words (e.g., *misunderstanding*, etc.) were associated with low concreteness and low valence responses (e.g., *argument*, *mistake*, etc.) and were generally less frequent than shorter words.

Overall, these examples and correlations indicate that there are indeed item-level influences of the cue upon R1. Furthermore, these ELP variables only showed low to moderate correlations with cosine similarities derived from word2vec ($r_{\text{R1-freq}}$ = -.25 , $r_{\text{R1-valence}}$ = -.07, $r_{\text{R1-concretness}}$ = -.16) and GloVe ($r_{\text{R1-freq}}$ = -.18 , $r_{\text{R1-valence}}$ = -.01, $r_{\text{R1-concretness}}$ = -.13), suggesting that these influences of cues' characteristics upon response production existed above and beyond the cosine similarity derived via DSMs. Of course, this may indicate that the ELP variables are capturing other non-linguistic aspects of semantic similarity or word meaning that are not effectively captured via DSMs, an issue that is discussed at length in the Discussion. To account for these item-level dependencies, the ELP model contained interaction terms for cue-R1 length, frequency, concreteness, and valence. As shown in Table 3.2, the ELP model accounted for 3.11% of the total variance in response probabilities in SWOW-R1, which was highly reliable, but much smaller than the amount of variance captured by the ELP model in the MEN (R^2 = 10.53%) and SimLex-999 (R^2 = 16.86%) datasets examined in Chapter 2. All subsequent models were incremental additions to the ELP model, to test whether the specific process or algorithmic models significantly improved model fit for R1 probabilities over the baseline ELP model.

Table 3.2. Explained Variance for R1 Probabilities in SWOW-R1

Structural Model	**Algorithmic/Process Model**	**SWOW-R1 Probabilities: R^2 [CI] (%)**
	ELP	3.11 [2.77, 3.39]
word2vec	Similarity*	8.92 [8.34, 9.44]
	Rotaru et al. model*	9.24 [8.63, 9.75]
	Luce*	9.77 [9.11, 10.36]
	SF$_{\text{additive}}$	4.23 [3.85, 4.57]
	SF$_{\text{multiplicative}}$	9.72 [9.07, 10.31]
	SF$_{\text{multiplicative}}$-delta*	11.90 [11.04, 12.66]
	SF$_{\text{multiplicative}}$-delta-neighbors*	**12.01 [11.12, 12.75]**
GloVe	Similarity*	9.13 [8.55, 9.65]
	Rotaru et al. model*	9.22 [8.62, 9.73]
	Luce*	9.55 [8.90, 10.15]
	SF$_{\text{additive}}$	4.51 [4.11, 4.85]
	SF$_{\text{multiplicative*}}$	9.68 [9.04, 10.26]
	SF$_{\text{multiplicative}}$-delta*	13.69 [12.76, 14.52]

	SF$_{\text{multiplicative}}$-delta-neighbors*	**14.00 [13.02, 14.85]**

Note: * indicates significant ($p < .05$) increase in variance based on log-likelihood tests over the previous model. CI indicates the 95% confidence interval based on bootstrapped R^2 estimates from 1000 samples with replacement.

As shown, the Similarity model explained significantly more variance than the ELP model, such that higher semantic similarities between the cue and R1 predicted greater likelihood of selecting R1. Furthermore, although variance increased in the Rotaru et al. model, compared to the ELP and the Similarity model, the Luce-choice model generally provided better model fits compared to the Rotaru model based on model likelihoods and confidence intervals, for both word2vec and GloVe. Next, although SF$_{\text{additive}}$ did not adequately explain R1 probabilities, SF$_{\text{multiplicative}}$ provided comparable fits to the Luce-choice model. It is important to reiterate here that the Luce-choice model was very similar in model formulation to the SF$_{\text{multiplicative}}$ model, therefore the comparable model fits were expected. Importantly, as shown in Table 3.2, in contrast to the results from Chapter 2, the delta models produced considerable increase in accounted variance compared to the other models. Thus, the competition between a response and competitor was quite powerful in predicting free association performance.

In order to further examine the nature of this pattern, Table 3.3 displays examples of when the cosine*frequency value in SF$_{\text{multiplicative}}$ (i.e., activation) of the specific competitor was high versus low, compared to R1 in the word2vec and GloVe-based SF$_{\text{multiplicative}}$ models.

Table 3.3. Examples of R1 probabilities against delta values in SWOW-R1

Structural Model	**Cue**	**R1**	**R1 probability**	**Competitor**	**Delta (R1- competitor)**
word2vec	cash	money	.95	buy	high
	village	town	.65	city	high
	ask	query	.05	know	low
	locate	GPS	.04	find	low
GloVe	elementary	school	.81	teacher	high
	hurricane	storm	.47	damage	high
	two	duo	.07	three	low

sort	arrange	.09	kind	low

As shown above, when this difference i.e., delta was high, the value of R1 in $\text{SF}_{\text{multiplicative}}$ was sufficiently higher than the next most active word, therefore overriding any competition to produce the specific response with high probability (e.g., producing the response *money* to *cash*). On the other hand, when delta was low, the value of R1 in $\text{SF}_{\text{multiplicative}}$ was not sufficiently high enough compared to the next most active word, leading to low overall likelihood to produce the response (e.g., producing the response *arrange* to *sort*). Indeed, it appears that there may be a *threshold* beyond which the difference between R1 and competitor values in $\text{SF}_{\text{multiplicative}}$ influences the likelihood of producing a particular response. In order to examine the relationship between delta and response probabilities for R1, Figure 3.2 plots delta against R1 probabilities, in the word2vec and GloVe $\text{SF}_{\text{multiplicative}}$ models. As shown, delta produced a highly reliable quadratic pattern ($p < .001$), and a significant positive correlation ($r = .26$, $p < .001$) with R1 probabilities, suggesting that there was a threshold beyond which the increased difference between R1 and competitor activation influenced the selection of that word as the final R1. Therefore, a quadratic term for *delta* was included as an additional predictor in the $\text{SF}_{\text{multiplicative}}$-delta regression model, to account for this process of comparing R1 responses to strong competitors. As indicated in Table 3.2, based on model likelihoods and confidence intervals, the $\text{SF}_{\text{multiplicative}}$-delta model consistently predicted more variance compared to the Luce-choice and $\text{SF}_{\text{multiplicative}}$ models ($p < .001$) in both word2vec and GloVe-based models.

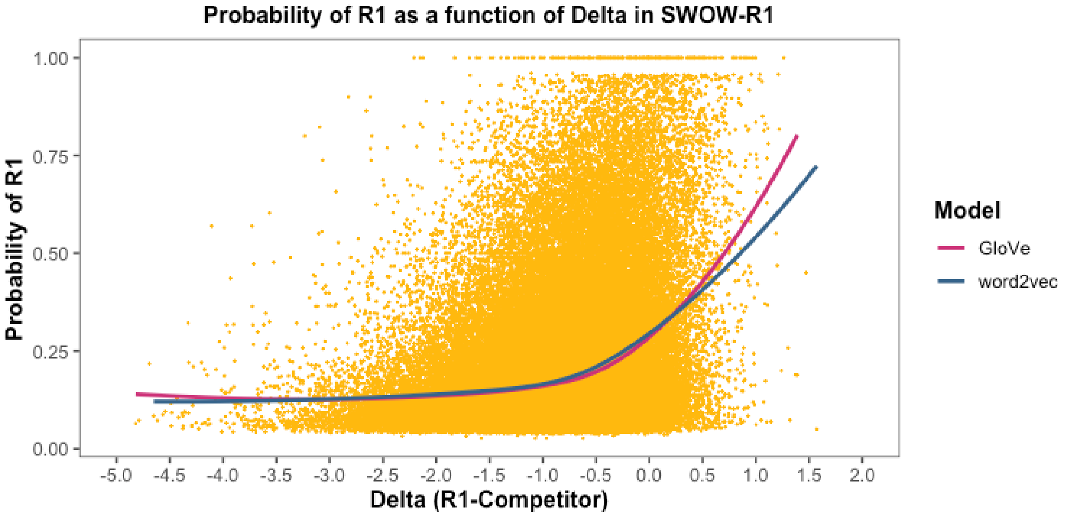


Figure 3.2. Probability of R1 as a function of delta (difference between R1 activations and competitor activations in $\mathrm{SF}_{\mathrm{multiplicative}}$) in SWOW-R1

Finally, there was a small but highly significant negative correlation (r = -.03, p < .001) between mean neighbor activations (indexed by values in $\mathrm{SF}_{\mathrm{multiplicative}}$) and R1 probabilities. As shown in Table 3.2, the $\mathrm{SF}_{\mathrm{multiplicative}}$-delta-neighbor model also improved overall variance explained in R1 responses, suggesting that the average level of competition within the network influenced response likelihoods. Thus, the best-fitting model was the ELP + $\mathrm{SF}_{\mathrm{multiplicative}}$-delta-neighbors model, based on model likelihoods and confidence intervals.

Figure 3.3 provides a way of conceptualizing the critical patterns obtained from the $\mathrm{SF}_{\mathrm{multiplicative}}$-delta-neighbors model, which shows the 3-way interaction between R1 values, delta, and the mean neighbor values in the $\mathrm{SF}_{\mathrm{multiplicative}}$-delta-neighbors model for the GloVe model[6]. As shown, the process of selecting a particular primary response involved higher activation of the specific response (influence by semantic similarity and frequency), greater difference between response activation and competitor activation, as well as low average

[6] Patterns were similar for the word2vec model

neighboring activations. Finally, as shown in Table 3.2, model fits were overall better for the GloVe model compared to the word2vec model. It is also important to highlight here that overall, explained variance was considerably lower in free association, compared to similarity/relatedness judgments in SimLex-999/MEN (Chapter 2), indicating that there were also systematic differences in the extent to which models derived from DSMs accounted for performance in production-based vs. familiarity-based tasks.

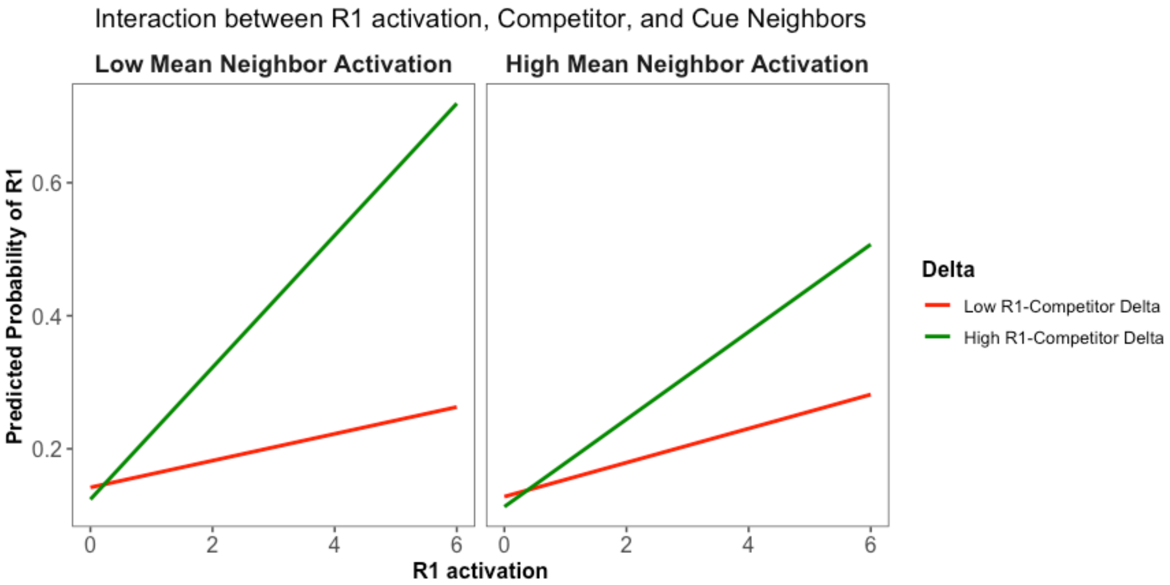


Figure 3.3. Predicted R1 probabilities as a function of R1 activation, delta, and mean neighbor activation in the SF$_{\text{multiplicative}}$-delta-neighbors GloVe model.

3.2.3 Predicting Primary (R1) Response Latencies

As noted, there has not been any work comparing specific models in predicting response latencies in free association. Hence, this section focuses on modeling the response latency data. Given that the responses were collected online and there were variable amounts of responses per participant, there was considerable variability in these data. Hence, in order to minimize the undue influence of extremely fast or slow RTs in the analyses, each individual's RTs were screened in the following manner for all analyses. First, RTs for R1 responses faster than 250 ms and slower than 5,000 ms were removed. This excluded 21.48% of the total trials, the majority of which (19.18% of total trials) were trials slower than 5,000 ms. This is not surprising, given that participants in the SWOW task were not encouraged to respond as fast as possible. These trials were therefore eliminated to ensure a quality dataset. It is important to note here that one can indeed use response latency data even when participants are not encouraged to respond quickly (see Aschenbrenner, Balota, Gordon, Ratcliff & Morris, 2016). Second, a mean and standard deviation were calculated from the remaining trials for each participant and any RTs that exceeded 3 standard deviations (SDs) from the participant mean were also removed (additional 1.07% of the remaining trials). Overall, these two screening steps excluded 22.3% of the total trials in SWOW-R1. After this trimming procedure, the remaining trials were standardized within each participant and all primary analyses were conducted using trial-level standardized RTs (z-RTs), to minimize any effects of general slowing and individual differences across participants (see Faust, Balota, Spieler, & Ferraro, 1999). As with response probabilities, the z-RTs for each unique cue-R1 combination were first aggregated, and the algorithmic and Rotaru et al. models were then applied to these aggregate estimates using linear mixed-effects models in R. A random intercept for the cue was included in all models to account for between-cue variability, and all algorithmic and process-level variables were included as fixed effects in the models. Table 3.4 displays the explained variance from the fixed and random effects in R1 z-RTs for the different structural and process models.

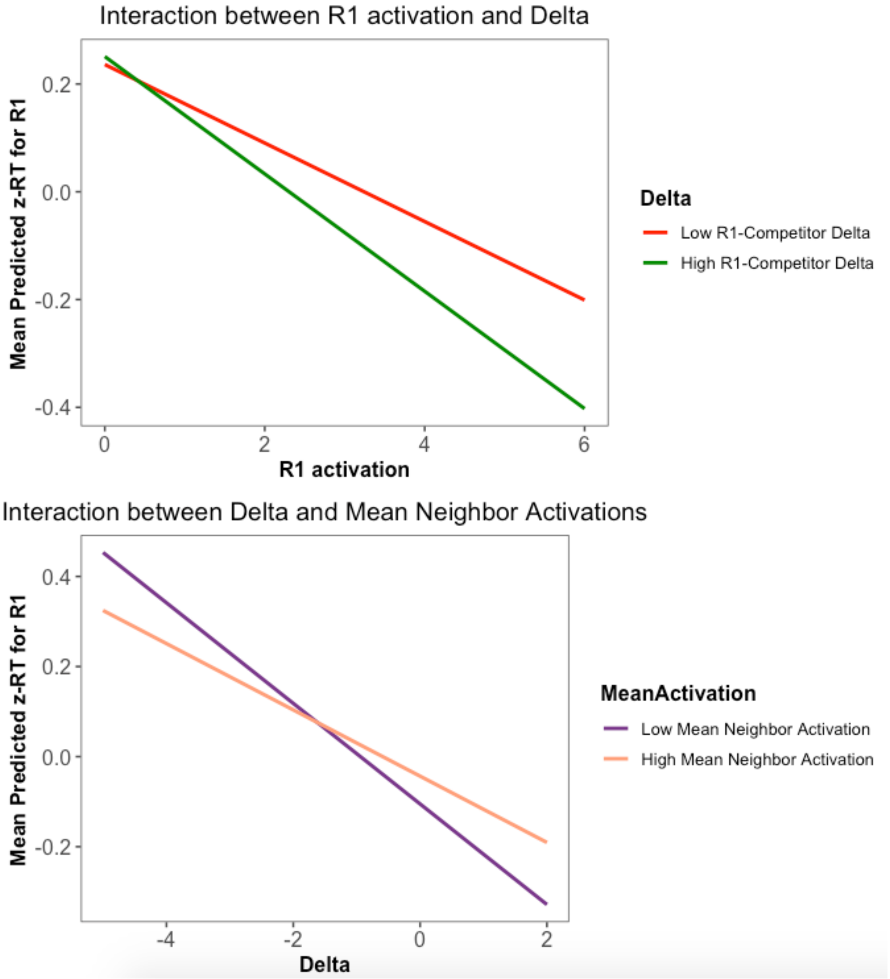
Table 3.4. Explained Variance for R1 z-RTs in SWOW-R1

Structural Model	**Algorithmic/Process Model**	**SWOW-R1 z-RTs: Fixed [CI] /Total R^2 (%)**
	ELP	3.12 [2.92, 3.56] /11.17
word2vec	Similarity*	4.15 [4.06, 4.81]/11.54
	Rotaru et al. model	4.18 [4.10, 4.85]/11.53
	Luce*	4.03 [3.96, 4.69]/11.47

	$\mathrm{SF}_{\mathrm{additive}}$	3.23 [3.05, 3.71] /11.20
	$\mathrm{SF}_{\mathrm{multiplicative}}$	4.04 [3.96, 4.70]/11.49
	$\mathrm{SF}_{\mathrm{multiplicative}}$-delta*	4.10 [4.05, 4.79]/11.49
	$\mathrm{SF}_{\mathrm{multiplicative}}$-delta-neighbors*	**4.19 [4.17, 4.91]/11.46**
GloVe	Similarity*	4.57 [4.56, 5.35] /11.62
	Rotaru et al. model	4.57 [4.55, 5.34]/11.62
	Luce*	4.40 [4.34, 5.10]/11.72
	$\mathrm{SF}_{\mathrm{additive}}$	3.30 [3.13, 3.79]/11.21
	$\mathrm{SF}_{\mathrm{multiplicative}}$	4.32 [4.32, 5.07]/11.47
	$\mathrm{SF}_{\mathrm{multiplicative}}$-delta*	4.38 [4.37, 5.12]/11.55
	$\mathrm{SF}_{\mathrm{multiplicative}}$-delta-neighbors*	**4.66 [4.64, 5.41]/11.70**

Note: * indicates significant ($p < .05$) increase in variance based on log-likelihood tests over the previous model. CI indicates the 95% confidence interval based on bootstrapped R^2 estimates from 1000 samples with replacement.

As shown, explained variance from fixed effects was overall low and confidence intervals for fixed effects overlapped across the models, indicating that the z-RT data was more difficult to fit, compared to response probabilities. This is a bit surprising, but likely reflects the considerable variation across participants, given that participants were not encouraged to respond quickly and on average only responded to 7 cues in SWOW-R1. Although the explained variance was relatively low, as shown in Table 3.4 the results were generally consistent with the response probability data. Specifically, the Similarity model explained significantly more variance compared to the ELP model. The Rotaru et al. model did not explain significantly more variance than the similarity model across both word2vec and GloVe models. Importantly, however, the $\mathrm{SF}_{\mathrm{multiplicative}}$-delta-neighbors model explained significantly more variance in fixed effects compared to all other models across both word2vec and GloVe, although confidence intervals overlapped across the models and the 3-way interaction was not significant for z-RTs. There were significant two-way interactions between R1 activation and delta, as well as delta and mean neighbor activations. As shown in Figure 3.4 (top panel), R1 responses were produced faster when the response was highly activated (as indexed by the value of R1 in $\mathrm{SF}_{\mathrm{multiplicative}}$) and the difference in activations between the response and competitor was high (as indicated by *delta* derived from SF$_{\text{multiplicative}}$). Furthermore, as shown in the bottom panel of Figure 3.4, when delta was high, lower mean activations produced faster z-RTs, whereas when delta was low, higher mean activations produced faster z-RTs. Finally, consistent with the primary response analyses, the GloVe model provided better model fits compared to the word2vec model.


Interaction between R1 activation and Delta

Interaction between Delta and Mean Neighbor Activations

Figure 3.4. Two-way interactions between R1 activation and delta, and delta and mean neighbor interactions in the SF$_{\text{multiplicative}}$-delta-neighbors GloVe model for R1 z-RTs.

3.2.4 R1 Rank Correlations

In order to further assess how well estimates from different models correlated with the general pattern of R1 responses for a given cue, *rank correlations* (Kendall's tau) between R1 and estimates from the similarity model, the Rotaru and Luce-choice models, and the SF models ($SF_{additive}$ and $SF_{multiplicative}$) were computed. For example, for a given cue *chair*, the ranking of different R1 responses was *sit* (1), *seat* (2), *table* (3), and *couch* (4), i.e., *sit* was produced most frequently, and *couch* was produced least frequently as the primary response. For each of these responses, estimates of "activation" (as indexed by cosine/frequency values within the different models) were computed and then these model estimates were correlated with the ranks of the actual R1 responses. For example, within the $SF_{multiplicative}$ model, the ranking of the responses based on the cosine*frequency values was *sit* (1), *seat* (2), *couch* (3), and *table* (4), and therefore the Kendall's tau rank correlation between the ranks based on actual probabilities vs. the $SF_{multiplicative}$ model estimates for this particular cue *chair* was $r = .67$. These rank correlations were computed for each unique cue within each model to ultimately obtain average rank correlations for each model. In this way, the rank correlations assessed which model best explained the general *pattern* of R1 responses produced in SWOW-R1.

Table 3.5 displays the rank correlations for the different models within the two structural DSMs for both R1 probabilities and z-RTs. As shown, the $SF_{multiplicative}$ and the Luce-choice model produced the highest rank correlations, which were significantly higher compared to all other models (p's < .05). Additionally, GloVe rank correlations were higher than word2vec rank correlations overall (p's < .05). It is important to note here that rank correlations of model estimates with z-RTs were modest ($r_{max} = -.07$), but this is not surprising given that R1 probabilities themselves were only moderately correlated with z-RTs ($r = -.18$, $p < .001$). Importantly, although rank correlations with z-RTs were small, they still showed patterns consistent with response probabilities, such that the SF$_{\text{multiplicative}}$ and the Luce-choice models produced the highest correlations.

Table 3.5. Rank Correlations in SWOW-R1

Structural Model	**Algorithmic/Process Model**	**Rank Correlation with Probabilities**	**Rank Correlation with z-RTs**
word2vec	Similarity	.17	-.03
	Rotaru et al. model	.15	-.03
	Luce	.23	-.07
	SF$_{\text{additive}}$	.12	-.07
	SF$_{\text{multiplicative}}$	**.23**	**-.07**
GloVe	Similarity	.18	-.04
	Rotaru et al. model	.16	-.03
	Luce	.24	-.07
	SF$_{\text{additive}}$	.12	-.07
	SF$_{\text{multiplicative}}$	**.24**	-.07

Of course, it is important to note here that the rank correlations above do not take delta and mean neighbor activations into account, given that delta and mean neighbor activations are derived from the SF$_{\text{multiplicative}}$ activation matrix and do not index the cue-R1 “activations” themselves. Indeed, as indicated by the regression models discussed earlier, delta and mean neighbor activations may reflect differentially weighted and potentially nonlinear relationships of each response against its competitors. Therefore, although we see identical correlations for Luce-choice and SF$_{\text{multiplicative}}$ (given their similar mathematical formulation), the regression models clearly indicate that the SF$_{\text{multiplicative}}$-delta-neighbors model best captures patterns in responses and z-RTs. Overall, despite predicting ranks equally well, taken together, the regression and rank correlational analyses provide converging support for the multiplicative delta-neighbors model over and above the Luce-choice model in accounting for responses and response latencies.

3.2.5 Comparing SWOW and USF norms

Given the interactions observed between R1 model estimates, delta, and mean neighbor estimates in the SWOW-R1 database, it is possible that these patterns were simply idiosyncratic to the SWOW-R1 database. Hence, it is important to extend these models to other free association norms. Specifically, these analyses compared the extent to which the same structural and process models accounted for patterns of free association in a subset of primary free association norms collected by Nelson et al., USF-4985 (see Methods section). First, it should be noted that R1 probabilities in USF-4985 and SWOW-4985 were strongly correlated (*r* = .74, *p* < .001). Next, as shown in Table 3.6, variance explained was slightly higher for USF-4985, but showed similar patterns, such that $\text{SF}_{\text{multiplicative}}$-delta-neighbors was still the best-performing model.

Table 3.6. Explained Variance for R1 probabilities in SWOW-4985 and USF-4985

Structural Model	**Algorithmic/Process Model**	**SWOW-4985 Probabilities: R^2 [CI] (%)**	**USF-4985 Probabilities: R^2 [CI] (%)**
	ELP	2.92 [2.46, 3.26]	3.08 [2.65, 3.43]
word2vec	Similarity*	10.24 [9.37, 10.99]	10.97 [10.18, 11.69]
	Rotaru et al. model*	10.50 [9.61, 11.27]	11.32 [10.50, 12.06]
	Luce*	10.84 [9.91, 11.69]	11.98 [11.12, 12.80]
	$\text{SF}_{\text{additive}}$	4.13 [3.59, 4.56]	4.17 [3.70, 4.58]
	$\text{SF}_{\text{multiplicative}}$	10.78 [9.85, 11.61]	11.70 [10.85, 12.50]
	$\text{SF}_{\text{multiplicative}}$-delta*	12.99 [11.75, 14.06]	14.14 [12.97, 15.22]
	$\text{SF}_{\text{multiplicative}}$-delta-neighbors*	13.10 [11.79, 14.12]	14.33 [13.10, 15.38]
GloVe	Similarity*	10.57 [9.76, 10.31]	11.21 [10.42, 11.93]
	Rotaru et al. model*	10.82 [9.86, 11.64]	11.55 [10.70, 12.30]
	Luce*	11.04 [10.12, 11.91]	12.19 [11.28, 13.04]
	$\text{SF}_{\text{additive}}$	4.39 [3.85, 4.84]	4.44 [3.95, 4.86]
	$\text{SF}_{\text{multiplicative}}$	11.05 [10.20, 11.85]	11.94 [11.10, 12.73]
	$\text{SF}_{\text{multiplicative}}$-delta*	15.26 [13.96, 16.48]	16.35 [15.06, 17.58]
	$\text{SF}_{\text{multiplicative}}$-delta-neighbors*	**15.78 [14.38, 17.00]**	**17.08 [15.68, 18.34]**

Note: * indicates significant (*p* < .05) increase in variance based on log-likelihood tests over the previous model. CI indicates the 95% confidence interval based on bootstrapped R^2 estimates from 1000 samples with replacement.

To further understand how well the different models captured the overall pattern of responses in the USF-4985 database, rank correlations of response probabilities in USF-4985 were computed for each algorithmic model and the Rotaru et al. model based on word2vec and GloVe. Given that the USF database does not contain RTs, rank correlations were only computed for primary responses. As shown in Table 3.7, rank correlations in USF-4985 showed a similar pattern to SWOW, such that the Luce-choice and $\text{SF}_{\text{multiplicative}}$ model again produced the highest correlations with response probabilities.

Table 3.7. Rank correlations in USF-4985

Algorithmic/Process Model	**Rank Correlation-GloVe**	**Rank Correlation-word2vec**
Similarity	.22	.21
Rotaru et al. model	.20	.18
Luce	.24	.23
$\text{SF}_{\text{additive}}$	.09	.08
$\text{SF}_{\text{multiplicative}}$	**.24**	**.23**

One important difference between the USF norms and SWOW-R1 is that within the USF norms, all cues were normed by 150 participants on average (ranging from 94 to 206), whereas the total number of participants for a given cue ranged from 4 to 150 in SWOW-R1 (*M* = 60.92, *SD* = 16.37). Therefore, it is possible that further reducing the SWOW dataset to include only cues that have been normed by a larger number of participants may lead to gains in predictive power within the current models. To investigate whether variance explained in R1 probabilities and z-RTs increased as a function of more stringent exclusion criteria, the SWOW-R1 dataset was systematically reduced as a function of total number of responses contributing to a specific cue, and the variance explained by the predictors in $\text{SF}_{\text{multiplicative}}$-delta-neighbors model for GloVe was estimated[7]. As shown in Figure 3.5, explained variance systematically increased as

[7] Patterns were similar for word2vec

the total number of responses for a given cue increased, although this also resulted in significant reductions in the data, as indicated by the legend and labels in Figure 3.5 (e.g., only 197 unique cue-R1 combinations had at least 100 responses for that specific cue in SWOW-R1). Overall, these analyses indicate that norming these data with greater number of participants per cue could lead to increases in predictive power in these models. Importantly, however, the major observation from this comparison is that the rank ordering of the models is replicated in a totally different dataset collected within lab (USF-4985), as opposed to online (SWOW-4985).

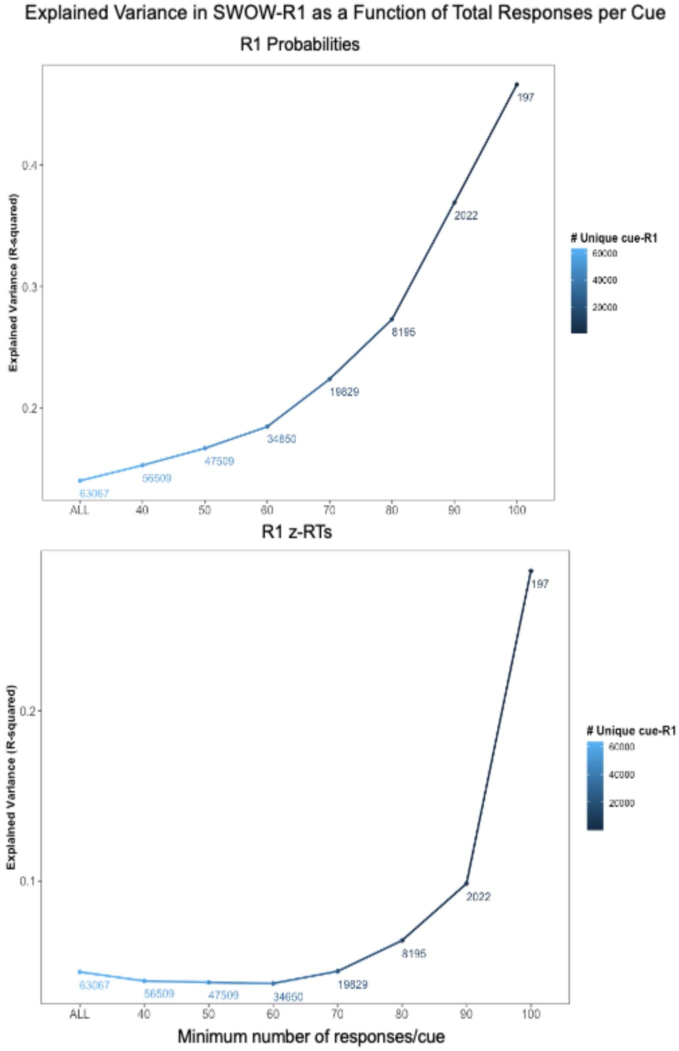


Figure 3.5. Percentage of explained variance in the $SF_{multiplicative}$-delta-neighbors GloVe model for R1 responses (top) and R1 z-RTs (bottom) as a function of minimum number of responses for a given cue in SWOW-R1. Numbers on curve indicate the total number of unique cue-R1 combinations within the reduced dataset.

3.2.6 Secondary Response (R2) Rank Correlations

Although the instructions in the SWOW task emphasized that participants should produce responses only to the cue item, it is possible that these instructions were not sufficient to eliminate any influence of the earlier primary response produced, since producing a specific R1 response is likely to place participants within a specific semantic space. Hence, it is likely that there will be some chaining of the responses, such that the second response will be influenced both by the cue and the first responses. Indeed, De Deyne et al. (2019) reported evidence for moderate chaining in the SWOW database based on contingency table analyses. However, the mechanisms by which such chaining might occur remain unclear.

To investigate the mechanisms influencing the secondary responses in the continued free association task, models for the secondary response parametrically explored whether additional activation from the cue vs. R1 within the $\mathrm{SF}_{\mathrm{multiplicative}}$ model-chained-cueR1 model influenced the ranks of R2 responses. The $\mathrm{SF}_{\mathrm{multiplicative}}$ model was chosen here because $\mathrm{SF}_{\mathrm{multiplicative}}$-delta-neighbors model consistently accounted for more variance in R1 response production, and the $\mathrm{SF}_{\mathrm{multiplicative}}$ model contains the underlying activations that drive the delta-neighbors model. For all R2 analyses, only trials on which at least 2 different R2 responses were produced to the same cue-R1 combination were considered, to effectively test whether the different models predicted one secondary response over another. Given that rank correlations were considerably informative in identifying best-performing models for R1, rank correlations were computed between the R2 probabilities and the value of different R2 responses within SWOW-R2 in the $\mathrm{SF}_{\mathrm{multiplicative}}$-chained models, in order to identify the best value of *theta* for a given DSM and neighbors above a certain number of standard deviations. Figure 3.6 shows the rank correlations between R2 probabilities and R2 activations, as a function of additional activation from the cue vs. R1 (in terms of *theta* and 1-*theta* respectively) and the neighbors (above 2 and 3 standard deviations) to which the activation was added within the SF$_{\text{multiplicative}}$ model-chained-cueR1 model based on GloVe and word2vec.

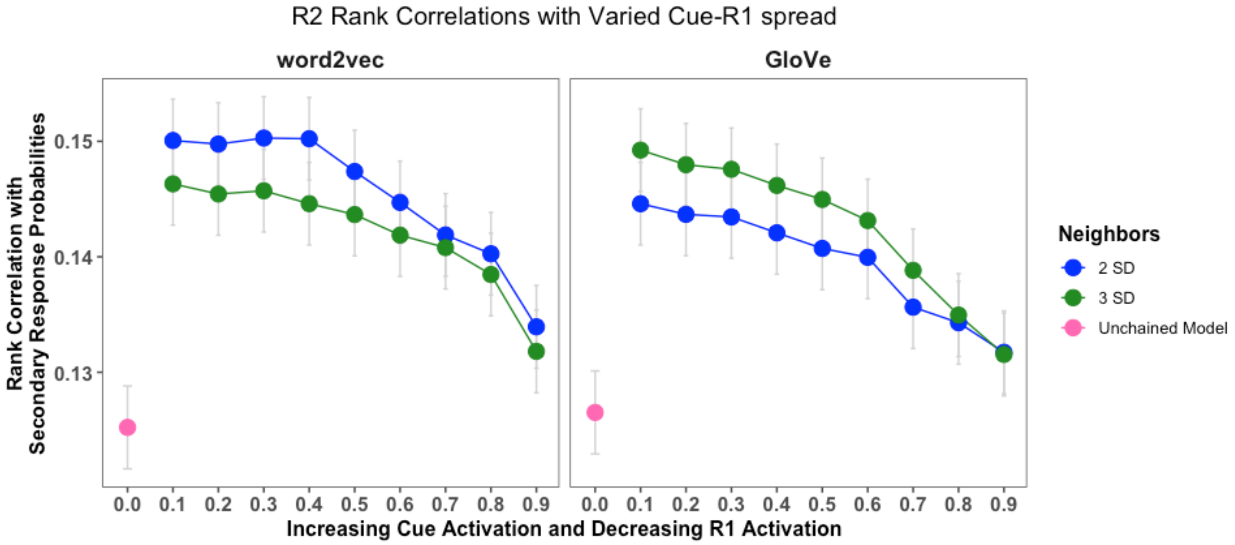


Figure 3.6. Rank correlations of R2 probabilities with R2 activations in models that simulated different amounts of activation spread from the cue vs. R1. Neighbors refer to semantic neighbors of the cue or R1 above a certain number of standard deviations (e.g., 2 or 3 standard deviations) to which activation was spread in the chained models. Error bars represent standard errors.

As is evident, increasing the relative activation from the cue to its neighbors led to significant *decreases* in rank correlations, whereas increasing the relative activation from R1 to its neighbors led to significant *increases* in rank correlations for *theta* greater than 0.4 (p's < .05), suggesting that R2 probabilities were best predicted when the additional activations to words in SF$_{\text{multiplicative}}$ were predominantly initiated from R1 to its neighbors. Moreover, rank correlations were overall higher for activations to neighbors above 2 standard deviations, compared to neighbors above 3 standard deviations (p's < .05). Indeed, the model that produced the highest average rank correlation with R2 probabilities was (0.3)*cue + (0.7)*R1 for the word2vec model (r = .15), and (0.1)*cue + (0.9)*R1 for the GloVe model (r = .149). These results indicate that the *chained* model (with increasing activation from R1, i.e., *theta* < .05) was more predictive of R2 responses compared to a model in which responses were primarily receiving additional activations from the cue (i.e., *theta* > 0.5) as well as an *unchained* model, in which no such additional activation was implemented (denoted via the pink points in Figure 3.6).

To further investigate how R1 influenced R2 responses, the *amount* of R1 activation (beta, ranging from 0.1 to 0.9) as well as the neighbors to which this activation was added (with similarities to R1 above n standard deviations, where n ranged from 1 to 5) was parametrically varied, as per the $SF_{multiplicative}$ model-chained-R1 model described in the Methods section. This resulted in 9 (beta) x 5 (n) = 45 model configurations for each structural DSM. To identify the best performing model among these 45 models, rank correlations were computed between the predicted ranks based on each $SF_{multiplicative}$ model-chained-R1 model as well as the original ranks based on R2 probabilities, for each unique cue-R1 combination. These correlations were then averaged to obtain a mean rank correlation estimate for each model configuration.

Figure 3.7 displays the mean rank correlations for different model configurations based on the $SF_{multiplicative}$ model-chained-R1 model. As shown, the model with 0.7 of R1's value in $SF_{multiplicative}$ being added to neighbors of R1 with similarity values over 2 standard deviations produced the highest correlations with subsequent R2 responses in the word2vec model (i.e., beta = 0.7, n = 2 standard deviations), whereas the model with 0.9 of R1's activation being added to neighbors of R1 with activations over 3 standard deviations produced the highest correlations with subsequent R2 responses in the GloVe model (i.e., beta = 0.9, n = 3 standard deviations). These best-fit parameters nicely converged with the theta-based parameters pertaining to the

SF$_{\text{multiplicative}}$ model-chained-cueR1 model reported above. Furthermore, models based on word2vec generally correlated higher with R2 probabilities, compared to models based on GloVe, in contrast to analyses based on R1 probabilities, where GloVe-based models outperformed word2vec-based models[8].

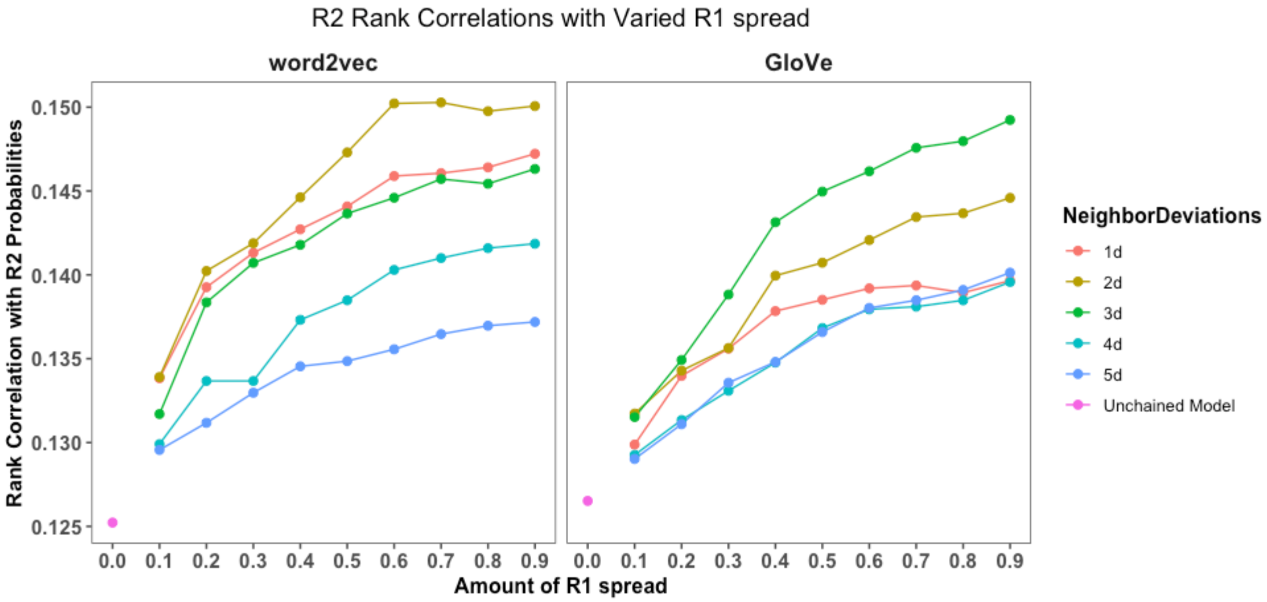


Figure 3.7. Rank correlations of R2 probabilities with models that simulated different amounts of R1 spread. NeighborDeviations refers to the semantic neighbors of R1 above a certain number of standard deviations (e.g., 2 standard deviations) to which activation from R1 was spread in the chaining model.

Although the chaining-based models with greater R1 spread better predicted R2 probabilities on average, there was considerable variability across different cue-R1 combinations. Table 3.8 displays examples of cue-R1 combinations where rank correlations were perfectly predicted by a chained model but not predicted by the unchained model (i.e., $r_{\text{chained}} = 1$ and $r_{\text{unchained}} = -1$; defined as "strong" chaining), as well as examples of cue-R1 combinations where the rank correlation was perfectly predicted by an unchained model but not predicted the chained model

[8] Rank correlations for R2 z-RTs were noisy overall, likely due to very few observations contributing to each cue-R1 combination and have therefore not been reported.

(i.e., $r_{\text{unchained}}$ = 1 and r_{chained} = -1; defined as “weak” chaining), within the word2vec and GloVe models. As shown, some cue-R1 combinations produced strongly chained responses with greater probability (e.g., *amazement-shock-awe*, *ink-blue-black*, etc.) whereas other combinations relied more heavily on the original cue and produced unchained responses with greater probability (e.g., *Aries-zodiac-goat*, *spirits-liquor-ghosts*, etc.). However, aggregating across all possible cue-R1 combinations, the chained model provided better fits to the R2 data, compared to the unchained model as indicated by rank correlation analyses above (and the subsequent regression analyses below).

Table 3.8. Examples of chained and unchained responses in SWOW-R2

Strong chaining (r_{chained} = 1, $r_{\text{unchained}}$ = -1)		
Structural Model	**Cue-R1**	**R2 responses (probability)**
word2vec	amazement-shock	awe (.67), happy (.33)
	popcorn-cinema	movie (.75), butter (.25)
	monkey-banana	tree (.67), ape (.33)
GloVe	clause-legal	lawyer (.75), Santa (.25)
	dim-dull	dark(.75), light (.25)
	ink-blue	black (.80), pen (.20)
Weak chaining (r_{chained} = -1, $r_{\text{unchained}}$ = 1)		
Structural Model	**Cue-R1**	**R2 responses (probability)**
word2vec	Aries-zodiac	goat(.67), horoscope (.33)
	nursing-mother	hospital (.67) baby (.33)
	right-correct	left (.86), wrong (.14)
GloVe	boa-feather	constrictor (.67), snake (.33)
	cow-animal	milk (.67), farm (.33)
	spirits- liquor	ghosts (.60), alcohol (.40)

3.2.7 Predicting Secondary (R2) Responses and z-RTs

The best-fitting models from SF$_{\text{multiplicative}}$ model-chained-R1 based on the rank correlations above (beta = 0.7, n = 2 for word2vec; beta = 0.9, n = 3 for GloVe) were used to further explore the influence of activation estimates from these models on the production of R2 responses and z-

RTs[9], for both GloVe and word2vec within regression models. Table 3.9 displays the contribution of R2 model estimates in the $\text{SF}_{\text{multiplicative}}$, $\text{SF}_{\text{multiplicative}}$-delta, and the $\text{SF}_{\text{multiplicative}}$-delta-neighbors models derived from the unchained and chained-R1 models in explaining variance in R2 probabilities and z-RTs[10]. As is evident, the $\text{SF}_{\text{multiplicative}}$-delta-neighbors *chaining-based* model was again the best-performing model within both GloVe and word2vec-based models based on model likelihoods and confidence intervals, and word2vec outperformed GloVe, consistent with the rank correlation analyses, although variance explained for R2 was relatively low compared to R1.

Table 3.9. Explained Variance for R2 probabilities and z-RTs in SWOW-R2

Structural Model	**Algorithmic/Process Model**	**SWOW-R2 Probabilities: Fixed [CI] /Total R^2(%)**	**SWOW-R2 z-RTs: Fixed [CI]/Total R^2(%)**
	ELP	1.03 [1.05, 1.27]/58.38	1.51 [1.48, 1.77]/23.07
word2vec-unchained	$\text{SF}_{\text{multiplicative}}$*	1.79 [1.84, 2.14]/58.95	1.49 [1.46, 1.75]/23.05
	$\text{SF}_{\text{multiplicative}}$ -delta*	2.94 [3.31, 3.71]/58.70	1.50 [1.48, 1.77]/23.04
	$\text{SF}_{\text{multiplicative}}$-delta-neighbors*	3.15 [3.82, 4.21]/57.58	1.54 [1.51, 1.81]/23.04
word2vec-chained-R1	$\text{SF}_{\text{multiplicative}}$*	1.58 [1.44, 1.74]/59.78	1.52 [1.49, 1.77]/23.07
	$\text{SF}_{\text{multiplicative}}$-delta*	1.60 [1.54, 1.82]/59.51	1.59 [1.56, 1.85]/23.06
	$\text{SF}_{\text{multiplicative}}$-delta-neighbors*	**4.96 [5.74, 6.23]/57.37**	**1.79[1.77, 2.07]/23.02**
GloVe-unchained	$\text{SF}_{\text{multiplicative}}$*	1.81 [1.83, 2.85]/59.02	1.48 [1.45, 1.74]/23.06
	$\text{SF}_{\text{multiplicative}}$-delta*	2.71 [2.96, 3.36]/58.72	1.51 [1.48, 1.77]/23.05
	$\text{SF}_{\text{multiplicative}}$-delta-neighbors*	3.18 [3.80, 4.21]/57.74	1.52 [1.50, 1.79]/23.04
GloVe-chained-R1	$\text{SF}_{\text{multiplicative}}$*	1.65 [1.53, 1.84]/59.73	1.48 [1.45,1.74]/23.06
	$\text{SF}_{\text{multiplicative}}$-delta*	1.84 [1.86, 2.17]/59.35	1.62 [1.61, 1.91]/22.98
	$\text{SF}_{\text{multiplicative}}$-delta-neighbors*	**3.68 [3.90, 4.33]/59.09**	**1.61 [1.59, 1.89]/23.01**

[9] Same exclusion criteria as before were applied to obtain R2 z-RTs, except that RTs greater than 8000 ms instead of 5000 ms were removed initially to account for potential slowing in second responses, which excluded 13.69% trials in SWOW-R2

[10] Other models (e.g., Rotaru, Luce, and $\text{SF}_{\text{additive}}$) were not examined for R2 given that $\text{SF}_{\text{multiplicative}}$-delta-neighbor models consistently provided better fits overall fits to the R1 responses.

Note: * indicates significant ($p < .05$) increase in variance based on log-likelihood tests over the previous model. CI indicates the 95% confidence interval based on bootstrapped R^2 estimates from 1000 samples with replacement.

3.3 Discussion

The results from the analyses revealed 6 major observations. First, the analyses indicated that there was some consistency in responses above and beyond association, with respect to simple item-level characteristics. Specifically, high frequency, high concreteness, and high valence cue words were more likely to produce responses with high-frequency, high concreteness, and high valence. These relationships appear to occur above and beyond the relationships captured by the DSMs, as indicated by the variance explained by the ELP model alone and low correlations between these variables and cosine similarities derived from word2vec and GloVe. Second, although there was a clear advantage of the process-based spreading activation model proposed by Rotaru et al. compared to the algorithmic models in accounting for variance in tasks that could be driven by the sum of activation processes, i.e., similarity and relatedness judgments (Chapter 2), this advantage of the Rotaru et al. process model was lost when the task was to select a candidate response from activated candidates, i.e., free association. Third, a multiplicative algorithmic model that incorporated semantic similarity and response frequency together with a delta function that computed the difference between the response and competitor “activations” ($SF_{multiplicative}$-delta), as well as a variable that accounted for mean level of neighbor activations for a given cue ($SF_{multiplicative}$-delta-neighbors model) best predicted free association responses and z-RTs. Fourth, a *chaining-based* model, that provided additional activation from the primary response to its neighbors best accounted for *secondary* responses and z-RTs, compared to an unchained model as well as models that provided an additional amount of activation to the cue’s neighbors. Finally, the GloVe-based algorithmic models better predicted *primary* responses and z-RTs, compared to word2vec-based algorithmic models, whereas this pattern was reversed for *secondary* responses and z-RTs. This section discusses implications for each of these findings in detail.

The power of the ELP variables in predicting R1 responses indicates that there are item-level biases in production-based tasks, such that item-level information about the cue tends to bias the lexical space even when the individual is not directly attending to this information during the task. It is possible that the semantic space itself may be biased towards capturing these lexical relations, but given the modest correlations of response frequency, valence, and concreteness with cosine similarities between the cue and response, the correlations in concreteness and valence may reflect the types of processing (i.e., mental imagery, emotional processing, etc.; De Deyne et al., 2021) that free associations tasks tend to evoke. Of course, there may be some shared variance between free association and the rating-based tasks used to obtain this concreteness/valence information. Given the lack of "pure" measures for such variables in the literature, the present work uses these ratings as an index of non-linguistic aspects of meaning. Future work should explore physiological and/or machine learning-based measures of emotion (e.g., Alm, Roth, & Sproat,2005; Westerink et al., 2008) and concreteness (e.g., Kounios & Holcomb, 1994; Lazaridou et al., 2015). Importantly, the present findings highlight how these non-linguistic relationships may be difficult to capture via distributional models trained solely on text corpora. Indeed, De Deyne et al. recently showed a sizeable advantage in using free association data to predict visual and affective feature norms, over distributional models based on linguistic corpora, as well as DSMs supplemented with additional feature-based information. The present findings converge with this work, suggesting that free associations do indeed reflect *multimodal* relationships, that may be difficult to capture from purely linguistic data that form the basis of distributional models tested in the present study. In addition, the length and frequency-based patterns were also particularly interesting, and likely reflect natural biases that individuals pick up on based on cue information within the free association task, i.e., when presented with uncommon (i.e., low frequency) long words, individuals are more likely to also generate similar types of words as responses. Importantly, this tendency to produce similar words may not be attentional or conscious, and may instead be driven by natural heuristics and biases (Kahneman, 2011). Overall, these patterns suggest that it is important to assess the contribution of item-level biases when accounting for performance within semantic tasks.

Importantly, however, in the free association task, the Rotaru et al. model was significantly outperformed by the multiplicative algorithmic model of semantic similarity and frequency, combined with additional variables that captured activations of strong competitors in predicting responses and z-RTs. This finding is critical, because it suggests that tasks that require selection of a single response from a pool of activated candidates within a semantic space requires different processing assumptions compared to tasks that are driven by overall activations within that space. Indeed, the attentional demands of selecting a single response are likely to be different from the task of judging similarity/relatedness between words, and the present study highlights how a process model based on overall activations that may be well suited to capturing similarity/relatedness judgments may not directly apply to a production-based task such as free association. It is also important to note here that the frequency-based Luce-choice model produced nearly identical patterns to the multiplicative model without delta and neighbor variables ($SF_{multiplicative}$; in responses, z-RTs, and rank correlations), which is consistent with their mathematical formulations (see Methods section). However, additional process-level assumptions about competitors and activated neighbors significantly improved model fit, suggesting that the process of producing free associations involves not merely selecting responses based on relative semantic similarity and frequency (as a frequency-based Luce-choice model would predict), but also attending to the *differences* in activation levels of surrounding activated words in memory. Indeed, the influence of this delta showed a quadratic pattern with response probabilities, suggestive of a threshold wherein differences in activations between a specific response and its strongest competitor beyond a specific value was sufficient to drive the response that was ultimately produced. Furthermore, this delta function not only predicted response likelihoods, but also predicted the time taken to produce a given response, such that greater differences between response and competitor activations led to faster responses.

Given that the models implemented above were based on data *aggregated* across participants, it may be the case that these effects do not reflect processes at play at the *participant* level. To investigate this issue, *trial-level* LME models with a random effect at the participant level, and a random slope for the delta-based predictor were implemented. These analyses again revealed a highly significant effect of delta (p <. 001) and normally distributed variation in the slopes for delta fitted at the participant level (M = -.006, SD=1.83 x 10^{-10}), suggesting that these effects were reliable even after accounting for individual-level variation within the predictive modeling approach. The General Discussion further elaborates on these issues pertaining to individual differences.

In addition to the effects of delta on response likelihoods and latencies, the average activations of neighbors beyond the strongest competitor also predicted responses and latencies, such that excessive competition (as indexed by high mean activations) predicted lower response likelihoods as well as slower responses. Finally, the algorithmic ($SF_{multiplicative}$-delta-neighbors) model also successfully explained the most variance in a completely different lab-based dataset of free associations, USF-4895, providing converging evidence in favor of this model. Taken together, these findings shed light on the competitive mechanisms underlying response selection and production within a continued free association task.

With respect to secondary responses, although instructions encouraged participants to focus on the cue, the act of producing a primary response is likely to situate an individual within a specific semantic context, which is likely to bias the secondary response. Consistent with this hypothesis, a chaining-based model that provided additional activation to semantic neighbors of the primary response best predicted secondary responses and z-RTs. Indeed, parametric analyses explicitly compared the contribution of the cue vs. the primary response within a chaining-based model, and found that greater activation from the cue, compared to the first response, actually decreased the predictive power of the models in capturing the overall pattern of secondary responses produced. This result is important as it highlights how an individual may generate successively dependent responses in a continued free association task. De Deyne et al. previously used contingency tables to show evidence for moderate chaining in the SWOW database, but did not explicitly model *how* this chaining might occur. Therefore, the present work presents a novel algorithmic account for chaining in continued free associations. Of course, as discussed earlier, there was variability within SWOW-R2 in the extent to which the different cue-R1 combinations showed clear chaining (see Table 3.8 for examples). Furthermore, although there was strong evidence for a chaining-based model (based on rank correlations, model likelihoods, and explained variance), it is important to note here that overall correlations of the chained model with R2 probabilities were moderate (r_{max} = .15), and explained variance was lower compared to primary responses. This suggests that the present algorithmic model is a first step in understanding the mechanisms by which subsequent responses are selected within a continued free association task. It is also important to note here that the present work only examined secondary responses, and it is possible that tertiary responses show even stronger evidence for chaining, which is an avenue for future work.

Finally, comparisons between the structural models (word2vec and GloVe) showed that GloVe generally outperformed word2vec in algorithmic and process-level models of primary free associations, but the pattern was reversed in secondary associations, such that word2vec better captured secondary associations. Why might one see opposite patterns in primary vs. secondary responses? Free association responses generally tend to reflect both similarity *and* relatedness, where similarity is often considered a special case of relatedness (De Deyne et al., 2019). Therefore, similar to the patterns observed in relatedness/similarity judgments, given that GloVe is more likely to capture different types of semantic relationships (that correspond to both relatedness *and* similarity) via co-occurrence ratios, compared to predicting words within a sentence as in word2vec (which may emphasize similarity more than relatedness), it follows that GloVe would better predict free associations (that capture both similarity and relatedness), compared to word2vec, although future work should focus on more controlled tests of these hypotheses.

Interestingly, as noted, the analyses also indicated that word2vec better captured *secondary* associations, compared to GloVe. Although this is surprising, it is possible that the nature of secondary responses is different from the nature of primary responses produced in the SWOW task. Indeed, De Deyne and Storms (2008) used Dutch continued word associations and showed that the pattern of taxonomic properties (i.e., the occurrence of responses that were superordinate, coordinate, synonyms etc.) varied across primary and secondary responses, such that the difference between the frequency of superordinate (e.g., *apple-fruit*) vs. coordinate (e.g., *apple-orange*) pairs was greater in primary responses (26% vs. 13%) compared to secondary responses (12.1% vs. 9.2%). Although it is unclear why word2vec might be better at capturing this information than GloVe, this study is useful in illustrating that primary and secondary responses may have different taxonomic distributions, which may in turn affect the extent to which different structural DSMs capture these responses. Future work should compare the taxonomic distribution of primary and secondary responses in the SWOW database as well as perform more focused comparisons between different structural models trained on the same corpora, to further clarify the locus of these structural model-based differences in accounting for free association performance.

In sum, the present study provided a novel approach to accounting for continued free associations based on a multiplicative model of semantic similarity derived from distributional models and frequency, and showed that the processing operations within a production-based attentional task such as free association systematically differ compared to a familiarity-based task such as providing similarity/relatedness judgments. Collectively, these results shed light on the dynamics of how responses are produced within a free association task, and show that response production in free association is a function of both overall activation level of a specific response as well as how these activation levels compare to other activated words in the semantic space.

Chapter 4: Modeling Cloze Responses and Latencies

Most individuals can effortlessly predict the end of a relatively constrained sentence. A common measure of this ability is the Cloze task (Taylor, 1953), where participants are presented with sentence fragments (e.g., “the amazing astronaut orbited the”) and asked to complete the fragment with the most likely next word (e.g., *planet*, *moon*, etc.). This task has been widely used in the language processing literature to study predictability and sentence comprehension (e.g., Rayner & Well, 1996; Sheridan & Reingold, 2012). However, the nature of the underlying semantic representations and processes used to access and produce the Cloze task response have not been thoroughly investigated. This study evaluates the extent to which different distributional models, when combined with appropriate algorithmic models predict both response proportions as well as RTs in the Cloze task. Importantly, although prior work has examined variables that influence Cloze responses (e.g., Smith & Levy, 2011) and latencies (Staub et al., 2015), no studies have investigated the extent to which a *distributional* model that learns semantic representations from text corpora predicts Cloze task performance. Therefore, this study will provide novel insights into how semantic information may be accessed and combined to produce responses in the Cloze task. The following sections briefly describe the specific distributional models (or “structural models”) and algorithmic models that will be tested within this study.

4.1 Structural DSMs

As discussed in Chapters 1 and 2, word-level distributional models such as word2vec and GloVe learn semantic representations from text corpora and represent word meaning in a high-dimensional vector space. These types of semantic representations likely provide some constraint during the retrieval of the final word in the Cloze task. For example, it is likely that semantic representations of words contained within a particular Cloze fragment specifically activate certain other words, which in turn influence the extent to which a response is activated and ultimately produced within the task. Therefore, as in the previous Chapter, word2vec and GloVe models will be investigated in conjunction with different algorithmic models to predict Cloze task performance.

In addition to the word-level embedding models (i.e., word2vec & GloVe), advancements in natural language processing and machine learning have led to the development of some newer models that use multi-word sentential context to derive a word's meaning. The underlying assumption in these models is that the meaning of a word strongly depends on the linguistic context (e.g., the word *bank* can have a financial and riverside-related meaning) within which the word is embedded, and words do not have "context-free" representations. In particular, the incorporation of an "attentional*"* component into the process of developing semantic representations has been a major breakthrough in this field. This component (Bahdanau et al., 2014) allows for attention to be focused on a subset of the original words within a sentence by increasing their weight based on positional and semantic information. Specifically, when encoding the representation of a given word (e.g., *bank*) in a sentence (e.g., "I went to the *bank* to withdraw money"), attention-based models assign different weights to all words in the sentence proportionate to their contribution (calculated via prediction error) in determining the meaning of the given word (e.g., *withdraw* and *money* would be weighted more than *went* when determining the representation of *bank*)[1]. Importantly, the specific weights or "attention scores" assigned to different words within a sentence vary depending on the task at hand - for example, the noun and verb may be critical in a sentence prediction task, whereas adjectives may be critical in a sentiment classification-type task. Of course, this notion of "attention" is metaphorical, and likely does not fully map onto the cognitive construct of attention and simply represents a way of quantifying how a machine learning model learns to adequately weight different parts of a sentence to improve its predictions.

Attention-based neural networks (NNs) are currently being widely applied to technologies like Google Translate and Siri, and form the underlying machinery of several state-of-the-art language models, such as Google's Transformer (Vaswani et al., 2017), BERT (Devlin, Chan, Lee, & Toutanova, 2019), OpenAI's GPT-2 and GPT-3 (Brown et al., 2020; Radford et al., 2019), and Facebook's RoBERTa (Liu et al., 2019). These models use multiple layers of "attention" and positional information to process words in parallel. For example, Google's BERT model is trained to predict the words hidden by a [mask] in a sentence (e.g., I went to the [mask] to buy a carton of milk; predict *store*) in the spirit of the Cloze task[2]. BERT computes probabilities for words that would fit the [mask] using the same implementation of "attention", i.e., by assigning different weights to different parts of the sentences within an error-driven machine learning framework. Importantly, the architecture of BERT allows it to be flexibly *finetuned* and applied to other semantic tasks, while still using the basic attention-based structure, where words within a sentence are differentially weighted within other words' vector representations based on their relative positions in the sentence and over several neural network

[1] See https://jalammar.github.io/illustrated-transformer/ for a detailed explanation of "attention" within neural networks

[2] BERT is also trained on an additional task of next sentence prediction, see Devlin et al. for details

layers and iterations. This framework turns out to be remarkably efficient and models based on the general Transformer architecture (e.g., BERT, RoBERTa, & GPT-2/3) outperform models that propose “context-free” semantic representations such as word2vec and GloVe (hereafter referred to as non-contextual models) on a battery of semantic tasks such as question answering, classification, and commonsense inference (Devlin et al., 2019). However, BERT has only recently been applied to a limited set of cognitive tasks such as predicting feature norms (see Bhatia & Richie, under review) and to my knowledge, has never been applied to account for response latency data.

Clearly, a contextual model like BERT is considerably different from non-contextual models such as word2vec and GloVe, given that the latter models do not incorporate mechanisms for assigning differential weighting to different parts of multi-word contexts in developing word representations. For instance, within word2vec and GloVe, the representation of a word such as *star* will be identical even if it is used in entirely different contexts, e.g., “The sun is a bright star” vs. “Tom Cruise is a Hollywood star”, whereas BERT will differentiate the vector representation across the two sentences based on the words surrounding *star* within a sentence. Therefore, within BERT, the word *star* will have a different vector representation for every sentence that it is part of, constructed using the attentional weighting mechanism discussed above. Tasks that involve explicit retrieval from semantic memory within a given context (e.g., free association and Cloze task) may differentially emphasize the role of sentential context, and importantly, it remains unknown how modern DSMs may predict the time course of response production in these tasks. Therefore, the present study evaluates two non-contextual DSMs (word2vec and GloVe) and one contextual DSM (BERT) in the extent to which they account for Cloze task performance. Specifically, the present work uses a version of the BERT model that is specifically finetuned to predict masked words in a sentence to evaluate its performance against human responses in the Cloze task. BERT was selected as the attention-based DSM to test within the present study because of its current popularity in machine learning as well as its specific training on the Cloze task. It is important to note here that although BERT can be finetuned for other tasks, such as question answering and even predicting relatedness/similarity, the present work uses a publicly available version of the model that is specifically trained to perform the Cloze task[3]. Furthermore, it is crucial to not only understand the contribution of the underlying semantic representations (derived via DSMs) but also how these representations are applied within an algorithmic modeling framework to select responses within the Cloze task. Therefore, the following section provides a brief overview of the specific algorithmic models that will be evaluated in this dissertation for the Cloze task. As we shall see, these models are based on similar principles developed to account for performance in Chapters 2 and 3.

4.2 Algorithmic Models for the Cloze Task

4.2.1 ELP Model

As in previous chapters, the ELP model examined the influence of item-level characteristics on Cloze responses and latencies. Specifically, item-level information (length, frequency, and concreteness for each content word (adjective, noun, and verb) as well as each unique response was extracted, and the ELP model examined the relationship between content words and response characteristics. As before, response frequency was dropped from the ELP model when it was already incorporated into the specific mathematical formulation (e.g., in the Luce-choice

[3] Indeed, the masked language model version of BERT used in the current study explained 52% variance in MEN and 21% variance in SimLex-999 (lower than the cosine similarity models based on word2vec and Glove, see Table 2.1), and 3.8% variance in SWOW-R1 (lower than the ELP model, see Table 3.1) suggesting that specific finetuning may be critical for BERT.

and SF-based models) and all subsequent models were incremental additions to the basic ELP model, which primarily accounts for item-level variables.

4.2.2 Unchained Additive Models

Given an underlying semantic space, one possible account for how Cloze responses are generated may be that as an individual encounters critical content words in the Cloze fragment (e.g., “the amazing astronaut orbited the”), neighbors of those content words (e.g., *amazing*, *astronaut*, and *orbited*) are systematically activated and their activations are summed. Ultimately, the response with the highest summed activations is selected by an individual. It is also possible that frequency biases the extent to which these initial neighbors are activated, which in turn influences the summed activations that emerge from this model. Importantly, the underlying “activations” themselves could be derived from different model formulations, such as one based on purely similarity, or a combination of similarity and frequency, or even a process-based model such as Rotaru et al., as seen in Chapter 3. Therefore, as in Chapters 2 and 3, the present study explores the extent to which activations derived from the similarity (S), Rotaru et al. model, Luce, and additive and multiplicative similarity-frequency (SF) models account for Cloze responses and latencies within an unchained additive algorithmic model of summed activations.

4.2.3 Chained Additive Models

An alternative account for how individuals perform the Cloze task is that instead of activating neighbors of incoming content words independently, these neighbors are in fact activated sequentially, therefore capturing the syntactic structure of the Cloze fragment. Consider, for example, the fragment, “the amazing astronaut orbited the”, where the critical content words are *amazing*, *astronaut*, and *orbited*. Moreover, consider two potential completion responses, “moon” and “earth”. Within an *unchained* model, all content words would activate their neighbors independently (in proportion to their semantic similarity to these words within a DSM), and ultimately the activations of *moon* and *earth* would be compared after summing the activation they receive from *amazing*, *astronaut*, and *orbited*. However, within a *chained* model, first, when *amazing* is activated, its neighbors would be activated in proportion to their similarity to *amazing*. Next, when *astronaut* is activated, it would further activate its own neighbors *among* the already activated neighbors of *amazing*, therefore accounting for the previous word in the fragment. Specifically, words are activated only relative to the already activated words, i.e., no additional words are independently activated by *astronaut*. Similarly, when *orbited* is activated, it would activate its neighbors among the already activated neighbors of *astronaut*. Finally, activations for *moon* and *earth* would be summed from each content word to ultimately select a response. In this way, the chained model accounts for conditional dependencies within the Cloze fragment and may provide a better account of performance in the Cloze task. Therefore, the present study compares the predictive power of the unchained and chained model, for the different “activation” matrices derived via the similarity, Rotaru et al., Luce, and SF-based models based on word2vec and GloVe. Furthermore, given that the BERT model used in the current study is explicitly trained to predict masked words in a sentence, one can directly obtain likelihood scores for different responses for a given Cloze fragment and examine whether these likelihood scores predict Cloze responses and latencies.

4.2.4 Delta and Neighbor-based Models

In addition to the activations corresponding to a specific response, it is also possible that the *difference* between a response and its next most active competitor, i.e., delta, influences the extent to which a response may be produced, as well as the time taken to produce a response. For example, it is possible that when a response is sufficiently activated beyond a threshold, it is more likely to be produced, and such responses are also produced faster. Therefore, similar to Chapter 3, the delta-based model within the present study tested whether delta influenced Cloze probabilities and latencies. Additionally, the delta-neighbors model tested whether the mean activations of neighbors also influenced task performance, over and above the contribution of delta from the strongest competitor. Furthermore, an important finding in the Staub et al. study was that Cloze responses were produced faster for more constraining fragments. Constraint was defined in terms of the modal response probability, which nearly perfectly correlated with the total number of unique responses to a given fragment. To explore whether the current models account for this behavioral pattern, activations from the delta model within the present study were also used to test whether responses were faster for greater delta at different cloze probabilities.

4.2.5 Weighted Sum Models

In addition to examining the chained and unchained models, it is possible that different components of the Cloze fragment differentially influence task performance. For example, within the fragment “the amazing astronaut orbited the”, one would expect that the noun (*astronaut*) and verb (*orbited*) are more critical than the adjective (*amazing*) to the response being produced. To test this hypothesis, the weighted-sum models parametrically varied the contribution of different components of the Cloze fragment (adjective, noun, and verb) within the best-performing chained and unchained models based on word2vec and GloVe, and identified which specific combination of weights assigned to nouns, verbs, and adjectives are most predictive of the final response and response latencies in the Cloze task. Given that the BERT model already produces likelihood scores for different responses for a given Cloze fragment, to mirror the parametric analyses, the present study also evaluated whether reducing the information provided to BERT during prediction (by truncating the fragments and systematically removing the adjective, noun, and verb; see Ettinger, 2020 for a similar approach) influenced its performance.

4.3 Overview

The present study seeks to provide a computationally driven account of how responses are selected and produced within the Cloze task. There were two different sets of analyses. The first set of analyses ("Predicting Cloze Responses and z-RTs") compared the extent to which combining structural DSMs with appropriate algorithmic models (chained/unchained, similarity/similarity-frequency/Luce/Rotaru based, and likelihood score-based) influenced response production. These analyses also examined the extent to which competitor activations (indexed via delta- and delta-neighbor models) predicted task performance, and also whether delta could account for the behavioral pattern in Cloze z-RTs. The second set of analyses ("Weighted Sum Models") parametrically evaluated the contribution of different components of the Cloze fragment in predicting Cloze responses and latencies.

4.4 Methods

4.4.1 Dataset

Sentence completion data was taken from Staub et al. (2015) for 338 sentence fragments (Experiment 2) with varying degrees of item constraint. This dataset includes trial-level responses from 40 participants with corresponding response latencies to produce the response (hereafter referred to as the CLOZE-338 dataset). All sentence fragments in CLOZE-338 contained five words, and used the same sentence structure (i.e., The ADJ NOUN VERB(+past) DET ____). The determiner (DET), which was the last word of the fragment was the definite article (i.e., *the*) in 240 fragments, and the remaining fragments used possessive pronouns (i.e., *her*, *his*, *its*, *theirs*, *Santa's*; 67 fragments), pronouns, or quantifiers (i.e., *them*, *many*, *some*; 31 fragments) as the last word.

4.4.2 Structural Models

The same pretrained 300-dimensional word2vec and GloVe models described in Chapters 2 and 3 were used as the non-contextual structural models. In addition, for BERT, the BERTforMaskedLM (BERT-large) model[4] introduced by Devlin et al. (2019), made available by HuggingFace (Wolf et al., 2018) was used. This pretrained BERT model had a vocabulary of 30,522 words and was trained on a large Wikipedia corpus (2.5 billion tokens) as well as an additional BooksCorpus (800 million tokens). Therefore, although all DSMs were trained on Wikipedia corpora, GloVe and BERT were also trained on additional corpora, which may provide an advantage for these models, an issue that is discussed at length in the General Discussion section. For word2vec and GloVe, word vector representations for each word in all sentence fragments were obtained, and algorithmic models were applied to these vector representations. For all computations involving word2vec and GloVe, a vector space of 12,373

[4] Devlin et al. released two versions of BERT (BERT-base and BERT-large) trained on different number of parameters, and BERT-large generally performs better than BERT-base

words was assumed, which contained the 11,906 words used in Chapter 2 to ensure comparability, in addition to 467 unique words present in CLOZE-338 not contained within the 11,906 words[5]. Within the BERT model, given that it is trained to predict masked words in a sentence within a vocabulary of 30,522 words, likelihood scores for different potential responses were directly obtained and all subsequent analyses were conducted based on these scores.

4.4.3 Algorithmic Models

Different algorithmic models based on the structural models described above were explored in the extent to which they accounted for performance in CLOZE-338.

Unchained Additive Model. As noted earlier, an unchained additive model simply estimated the cosine similarity of critical content words in the fragment to potential responses. Given that the sentence fragments in CLOZE-338 were structured the same way and used a determiner/pronoun as the first and last (fifth) word in the fragment, only the second (ADJ), third (NOUN), and fourth (VERB) words were considered as "content" words[6]. Next, the unchained model also explored whether similarity alone, or other conceptualizations of the underlying activation matrix (Rotaru et al. model, Luce, and a similarity-frequency models) could account for performance. Specifically, the unchained model computed the sum of activations, between each content word and every possible response in CLOZE-338. For example, for the fragment, "the amazing astronaut orbited the":

$$\begin{aligned}\text{Sum-M}_{\text{unchained}} = {} & \text{M } (\textit{amazing}, \text{ response}) + \\ & \text{M } (\textit{astronaut}, \text{ response}) +\end{aligned}$$

[5] These words were mostly past-tense forms of verbs already contained within the 11,906 words, and were often the fourth word in the Cloze fragment such as *climbed*, *undermined*, *witnessed*, etc.

[6] Preliminary analyses showed that adding the fourth content word (which was the definite article in 70% of the fragments) did not change overall patterns

M (*orbited*, response), and

were computed for all responses in CLOZE-338 to this particular fragment, where M denoted the activation matrix derived from S/Rotaru/Luce/SF$_{additive}$ /SF$_{multiplicative}$, as described in earlier chapters. These activation estimates were then used in a regression model to test whether they accounted for Cloze probabilities and RTs in CLOZE-338.

Chained Additive Model. The chained model assumed that activation spread to neighbors of content words (as in the unchained model), but this spread was conditional on the previously activated content words, similar to the chained model described in Chapter 3. Specifically, the chained model was implemented as follows:

1. When the first content word (C1) was activated, it activated other words in the semantic space in proportion to their activation with respect to C1 (as indicated by S(C1, word), SF$_{multiplicative}$ (C1, word), Luce (word |C1) etc.) within the specific activation matrix.

2. When the second content word (C2) was activated, it added some amount of its activation to other words in the semantic space in proportion to the previous activations from C1. This process was implemented by adding a proportion (theta, parametrically varied) of C2's values in S/Rotaru/Luce/SF$_{additive}$ /SF$_{multiplicative}$ to a specific number of C2's neighbors (with cosine similarities to C2 over 3 standard deviations[7]). Therefore, when C2 was activated, it further activated some fraction of words that had already been activated due to C1. The same procedure was followed for C3 to ultimately yield *chained* activations within the specific activation matrix.

[7] Analyses with neighbors with similarity values over 2 SD resulted in lower overall explained variance

3. After obtaining these chained estimates of similarity/similarity-frequency, a sum was computed for activations between each content word (i.e., C1, C2, and C3) and each unique response produced by the participant for the fragment as follows:

$$\begin{aligned} \text{Sum-M}_{\text{chained}} = {} & \text{M}_{\text{chained}}\ (\textit{amazing}, \text{ response}) + \\ & \text{M}_{\text{chained}}\ (\textit{astronaut}, \text{ response}) + \\ & \text{M}_{\text{chained}}\ (\textit{orbited}, \text{ response}), \text{ and} \end{aligned}$$

Therefore, within a chained additive model, if a word (e.g., *planet*) was a neighbor of each of the content words, it would have a higher value in the underlying activation matrix M (i.e., S/Rotaru/Luce/$\text{SF}_{\text{additive}}$ /$\text{SF}_{\text{multiplicative}}$), compared to a word that received activation only from the first word (e.g., *wonderful*), which in turn may predict the extent to which that word may be selected as a response in the Cloze task. Note that the model formulation was identical in the unchained and chained models, with the exception of the underlying activations being derived directly from M in the unchained model, compared to $\text{M}_{\text{chained}}$ in the chained model.

Multiplicative Delta-Neighbor Models. As noted, it is possible that in addition to semantic similarities and frequency (as indexed by values $\text{SF}_{\text{multiplicative}}$), strong competitors may also influence the decision process of selecting a response in the Cloze task. To evaluate this possibility, the *difference* between the activations of the specific response and the next most active competitor for a given fragment (i.e., delta, as in previous chapters) within the $\text{SF}_{\text{multiplicative}}$ model was computed. Within the non-contextual DSMs (word2vec and GloVe), the strongest competitor was identified based on the words with the highest value in $\text{SF}_{\text{multiplicative}}$ with respect to the content words. Within BERT, the strongest competitor was identified as the next best completion predicted by the BERT model, other than the response itself. Therefore, the delta-based model examined how delta influenced the final response and RTs for the specific Cloze response, and whether delta could account for specific behavioral patterns in Cloze response latencies. Importantly, as in Chapters 2 and 3, a quadratic term for delta was added to regression models if delta showed a significant quadratic trend against Cloze probabilities or latencies. Finally, in addition to only examining the contribution of the strongest competitor, similar to Chapters 2 and 3 where the mean neighbor activations were examined, the delta-neighbors model evaluated how mean activations of *n* (fixed to 10)[8] neighbors predicted task performance.

Weighted-Sum Models. Finally, as noted, in addition to examining the predictive power of unchained and chained models in predicting Cloze task performance, the weighted-sum model parametrically varied the contribution of the different content words (C1, C2, and C3) to explore whether different parts of the sentence fragment (i.e., ADJ, NOUN, and VERB) were differentially influencing the extent to which a particular response was selected. Specifically, a "weighted sum" for the unchained and chained models was computed, such that

$$\begin{aligned}\text{Weighted-Sum-M} = {} & \alpha^{*}\text{M (C1, response)} + \\ & \beta^{*}\text{M (C2, response)} + \\ & \gamma^{*}\ \text{M (C3, response)},\end{aligned}$$

where M referred to the specific model being tested (similarity vs. similarity-frequency, and chained vs. unchained), and α, β, and γ denoted parametrically varied weights for C1, C2, C3, exploring all possible triplet permutations of weights in the range of 0.1 to 0.9 that summed to 1 (e.g., 0.1-0.1-0.8, 0.1-0.2-0.7, etc.), yielding a total of 36 unique combinations for each structural DSM and specific algorithmic model. In this way, the weighted-sum model assessed which specific combination of weights best predicted Cloze responses and latencies.

[8] Initial analyses examined a range of neighbors (i.e., n = 10, 20, 30, etc.) and n=10 produced best results

Given that the BERT model directly provides likelihood scores for different potential responses to a fragment, in order to assess how different content words contributed to BERT's predictions, words within the Cloze fragment were incrementally removed and likelihood scores for different responses were obtained for these truncated fragments. For example, for the complete fragment "the amazing astronaut orbited the", the truncated sub-fragments "amazing astronaut orbited the", "astronaut orbited the", "orbited the", and "the" were tested and the extent to which likelihood scores generated by the BERT model predicted Cloze responses and z-RTs was evaluated (see Ettinger, 2020 for a similar approach of truncating sentences provided to BERT). In this way, the truncated models assessed whether attending to specific parts of the Cloze fragment was more or less beneficial to predicting Cloze responses within the BERT model.

Overall, the present study evaluated the extent to which different structural DSMs and algorithmic models accounted for Cloze task performance, as well as whether additional assumptions regarding competitor activations and differential weighting of content words improved the predictive power of the chained/unchained based on word2vec and GloVe, as well as the BERT model based on likelihood scores.

4.5 Results

4.5.1 Predicting Cloze Responses and Latencies

To ensure that RTs were not influenced by outliers and individual differences, RTs above 2500 ms and below 250 ms were trimmed and then standardized RTs using the same procedure as in Chapter 3. This procedure excluded 2.47% of the total trials. All analyses were conducted on trial-level standardized reaction times (z-RTs), which were subsequently aggregated at the fragment level to obtain mean Cloze probabilities and mean z-RTs for each unique response.

Table 4.1 displays the explained variance in Cloze probabilities and z-RTs for the different structural and algorithmic models. The first set of analyses evaluated the extent to which ELP variables predicted Cloze responses and latencies. As shown, basic item-level characteristics of the content words within the ELP model strongly predicted Cloze responses and z-RTs. Specifically, responses with high concreteness, high frequency, and shorter lengths were more produced with greater probability overall (e.g., *dog*, *car*, *cat*, *bat*, etc.), although there were no significant interactions between response and content-word characteristics (*p*'s > .05).

The second set of analyses evaluated the extent to which unchained vs. chained models as well as the BERT model predicted Cloze responses and latencies. Theta was parametrically varied from 0 to 1 within the chained models (reflecting how much of C2 and C3's values in S/SF models was to be added to their neighbors) to obtain best-fitting model estimates within each structural DSM, and theta = 1 produced the best model fits across all DSMs.

Table 4.1. Explained variance in Cloze response probabilities and z-RTs

Structural Model	**Process/Algorithmic Model**	**Cloze Probability: Fixed [CI]/Total-R^2 (%)**	**Cloze z-RTs: Fixed [CI]/Total-R^2 (%)**
	ELP	11.10 [4.19, 13.32]/21.68	5.85 [1.91, 6.37]/14.65
word2vec unchained	Similarity	12.64 [6.17, 15.23]/21.31	6.00 [2.06, 6.50]/14.42
	Rotaru et al.	12.79 [6.24, 15.27]/20.41	6.25 [2.14, 6.70]/14.63
	Luce	12.45 [6.19, 15.22]/22.40	5.54 [1.90, 6.27]/13.98
	SF$_{\text{additive}}$	11.09 [4.42, 13.47]/22.24	5.47 [1.80, 6.18]/14.14
	SF$_{\text{multiplicative}}$	12.31 [6.03, 15.07]/22.34	5.55 [1.91, 6.28]/14.03
	SF$_{\text{multiplicative}}$-delta	12.55 [6.12, 15.16]/23.16	6.04 [2.30, 6.82]/13.95
	SF$_{\text{multiplicative}}$-delta-neighbors	**12.98 [6.38, 15.46]/23.64**	**6.20 [2.25, 6.85]/14.04**
chained	Similarity	12.60 [6.00, 15.04]/20.23	5.95 [2.00, 6.44]/14.48
	Rotaru	12.70 [6.04, 15.09]/20.22	6.10 [1.95, 6.49]/14.71
	Luce	12.39 [6.02, 15.03]/21.61	5.51 [1.87, 6.24]/13.99

	$\text{SF}_{\text{additive}}$	11.27 [4.60, 13.63]/21.46	5.48 [1.81, 6.20]/14.11
	$\text{SF}_{\text{multiplicative}}$	12.30 [5.93, 14.94]/21.45	5.52 [1.88, 6.25]/14.03
	$\text{SF}_{\text{multiplicative}}$-delta	12.47 [5.75, 15.07]/22.22	6.72 [2.63, 7.45]/14.35
	$\text{SF}_{\text{multiplicative}}$-delta-neighbors	12.53 [5.38, 14.70]/21.51	7.33 [2.74, 7.67]/15.64
GloVe unchained	Similarity	13.67 [7.57, 16.59]/23.43	5.91 [2.13, 6.56]/13.86
	Rotaru	14.03 [7.71, 16.77]/21.91	6.54 [2.45, 7.01]/14.71
	Luce	13.80 [7.96, 16.97]/22.43	5.84 [2.24, 6.64]/13.95
	$\text{SF}_{\text{additive}}$	11.11 [4.46, 13.51]/22.26	5.48 [1.81, 6.20]/14.15
	$\text{SF}_{\text{multiplicative}}$	12.54 [6.36, 15.37]/23/27	5.56 [1.98, 6.33]/13.78
	$\text{SF}_{\text{multiplicative}}$-delta	12.66 [6.15, 15.28]/23.96	6.06 [2.29, 6.79]/14.61
	$\text{SF}_{\text{multiplicative}}$-delta-neighbors	13.92 [7.57,16.75]/24.06	6.86 [2.90, 7.57]/14.63
chained	Similarity	14.01 [7.92, 16.94]/23.44	6.01[2.07, 6.54]/14.44
	Rotaru et al.	14.25 [7.8, 17.08]/23.01	6.37 [2.30, 6.85]/14.53
	Luce	14.07 [8.82, 17.28]/22.95	5.74 [2.15, 6.54]/13.85
	$\text{SF}_{\text{additive}}$	11.73 [5.25, 14.28]/21.80	5.52 [1.87, 6.26]/14.05
	$\text{SF}_{\text{multiplicative}}$	12.90 [6.78, 15.79]/23.27	5.57 [1.91, 6.30]/14.04
	$\text{SF}_{\text{multiplicative}}$-delta	13.75 [7.68, 16.94]/24.28	6.01 [2.20, 6.73]/14.55
	$\text{SF}_{\text{multiplicative}}$-delta-neighbors	**15.01 [8.68, 18.20]/25.79**	**6.47 [2.41, 7.05]/14.84**
BERT	Likelihood Scores	29.91 [23.93, 41.63]	7.41 [3.68, 8.08]/15.36
	BERT-delta	31.45 [24.26, 41.76]	8.56 [4.48, 9.08]/15.78
	BERT-delta-neighbors	**32.30 [24.48, 42.05]**	**9.65 [5.24, 9.96]/16.05**

Note: CI indicates the 95% confidence interval based on bootstrapped R^2 estimates from 1000 samples with replacement.

Within the non-contextual models (i.e., word2vec and GloVe), the GloVe model generally outperformed word2vec based on model likelihoods, and the *chained* GloVe model explained more variance than the *unchained* GloVe model based on model likelihoods and confidence intervals. However, within the word2vec model, the unchained model explained slightly more variance than the chained model, although confidence intervals largely overlapped across these estimates. Importantly, BERT substantially outperformed both word2vec and GloVe in predicting both Cloze response probabilities and z-RTs solely based on likelihood scores for different potential responses. However, it is important to note here that variance explained in z-RTs was low overall and the unchained word2vec and GloVe-based models performed relatively well in explaining z-RTs, compared to BERT, as is indicated by the confidence intervals around the R^2 estimates for z-RTs. Overall, however, the BERT-delta-neighbors model explained the most variance in responses and z-RTs, suggesting that accounting for competitor activations was critical in accounting for Cloze task performance. This nicely replicates the pattern observed in the free association responses in Chapter 3.

To demonstrate the influence of delta on Cloze responses, Table 4.2 displays some examples of delta-based competitors for different fragments within the different DSMs. As shown, when the difference between the next most active competitor and the possible response, i.e., delta was high, there was a greater likelihood of selecting that response (e.g., *hair* vs. *beard*), whereas when delta was low, the likelihood of selecting that response was also low (*lightbulb* vs. *coordinator*).

Table 4.2. Examples of Cloze probabilities against high vs. low delta

Structural Model	**Fragment**	**Response (probability)**	**Competitor**	**Delta**
word2vec	The pastry chef decorated the	cake (.78)	dessert	high
	The assistant manager replaced the	lightbulb (.03)	coordinator	low
GloVe	The reliable pilot landed the	plane (.91)	flight	high
	The crackling radio broadcast the	radio show (.03)	talk	low
BERT	The male model combed his	hair (.97)	beard	high
	The school lunch included the	cookie (.03)	following	low

Furthermore, Figure 4.1 displays the relationship of delta with Cloze probabilities and z-RTs within the different DSMs.

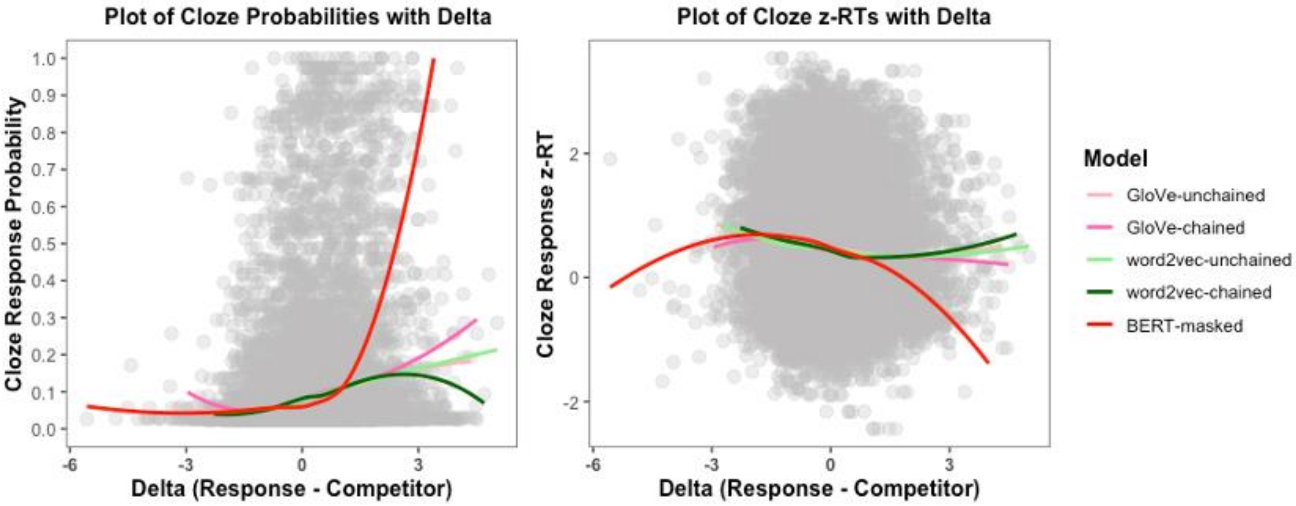


Figure 4.1. Cloze probabilities as a function of delta within different DSMs.

As shown in the left panel of Figure 4.1, delta showed a significant quadratic pattern with Cloze probabilities in the GloVe and BERT models (*p*'s < .05), indicative of a threshold, such that when response activations were sufficiently higher than competitor activations, the likelihood of producing that response increased. Within z-RTs, BERT showed a strong quadratic trend (*p* < .001) and word2vec showed a small but significant quadratic trend (*p* < .001), such that higher delta after a threshold led to faster responses beyond a particular threshold. Indeed, as shown in Table 4.1, the BERT-delta-model significantly predicted Cloze responses and latencies, and additional information about mean neighbor activations also improved variance estimates, such that greater mean neighbor activations facilitated response production and z-RTs.

An important observation from Staub et al. (2015) was that higher probability responses were produced faster and responses were produced faster in more constraining contexts. To explore whether the present computational framework of response and delta-based activations could account for this pattern in z-RTs, a regression model predicting z-RTs was implemented with fixed effects for cloze probability, delta, and an interaction term between the two, to account for the effect of cloze probability on z-RTs. These analyses again showed significant effects of delta on z-RTs, even after accounting for cloze probabilities, in word2vec, GloVe, and BERT models, and BERT still explained the most variance in this task. To better visualize the relationship between cloze probability, delta, and z-RTs, Figure 4.2 displays the time taken to produce a response (z-RT) as a function of Cloze probability and delta (right panel)[9], in comparison to the raw data (left panel).

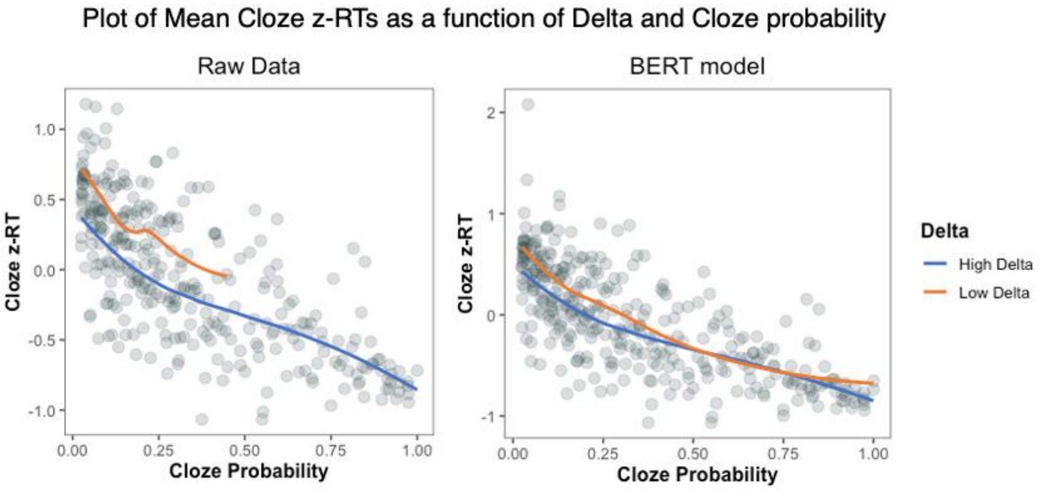


Figure 4.2. Mean z-RTs to produce a response in the Cloze task as a function of delta and Cloze probability. Delta within the raw data was defined through a median split on the total number of unique responses to a given fragment.

Indeed, as shown in Figure 4.2, delta predicted z-RTs even after accounting for cloze probabilities (i.e., there was a significant main effect of delta within the regression models, $p <$ .05), such that responses were indeed faster when delta was high and largely mirrored the relationship in the raw data based on high vs. low number of unique responses to different cloze fragments, therefore explaining a critical finding within the Cloze task.

[9] Patterns are displayed only for the BERT model, given that it was the best-performing model. High vs. low delta was defined based on median splits for delta within BERT.

4.5.2 Weighted Sum Models

In addition to exploring the contribution of chained and unchained models through the additive models above, the contribution of the adjective, noun, and verb was parametrically varied within the chained and unchained model sums, to investigate whether specific types of syntactic constraint within the Cloze prompt was guiding behavior in this task. After obtaining estimates of the weighted sums for all 36 triplet permutations (as described in the Methods section), linear mixed effects models predicting Cloze probabilities were implemented and the fixed and random R^2 explained by each possible model were estimated. Note that these analyses have been reported only for the delta-neighbors models for all DSMs, given that this model consistently outperformed other models. Further, given that the BERT model directly provided likelihood scores for different Cloze responses, the relative contribution of different content words within BERT was assessed by systematically removing words from the fragment provided to BERT (as described in the Methods section). Table 4.3 displays the weighted sum combination that produced the *highest* fixed R^2 for the best-performing model based on delta-neighbors models predicting Cloze probabilities, as well as the explained variance as the content words within the fragment were systematically removed within the BERT model.

Table 4.3. Explained variance in Cloze probabilities in weighted-sum models

Structural Model	**Chained/ Unchained**	**Cloze Probability: Highest Fixed R^2 [CI]/Total-R^2 (%)**	**Best Weighted Sum Combination/ Fragment (Adjective-Noun-Verb)**
ELP	-	11.10 [4.19, 13.32]/21.68	-
word2vec	Unchained*	13.55 [7.24, 16.37]/24.12	0.1+0.3+0.6
	Chained	13.50 [6.72, 16.27]/23.57	0.1+0.3+0.6
GloVe	Unchained*	16.37 [9.69, 20.09]/28.19	0.1+0.3+0.6
	Chained	15.36 [8.94, 18.51]/27.25	0.1+0.3+0.6
BERT*		32.30 [24.48, 42.05]	The adj-noun-verb-the

		34.64 [27.01, 45.28]	**adj-noun-verb-the**
		31.95 [24.59, 42.31]	noun-verb-the
		19.92 [11.54, 24.43]	verb-the
		12.05 [5.45, 14.60]	the

Note: * indicates significant (p < .05) increase in variance based on log-likelihood tests over the previous model. CI indicates the 95% confidence interval based on bootstrapped R^2 estimates from 1000 samples with replacement.

Within the non-contextual DSMs, the specific combination of adjective-noun-verb weights that produced the highest explained variance was 0.1 (adjective) + 0.3 (noun) + 0.6 (verb) across both word2vec and GloVe. Therefore, the verb and noun both contributed more than the adjective, but the verb was more critical than the noun in predicting Cloze response probabilities. Furthermore, GloVe again outperformed word2vec. Interestingly, within the BERT model, maximum variance was explained when BERT had access to the fragment without the definite article, and this model explained significantly more variance than the full-fragment model (*p* <. 001). Variance also significantly decreased when the adjective, noun, and verb were systematically removed (*p*’s < .05). As is evident, however, the most significant drops in variance occurred when the noun and verb were removed within the BERT model, which is consistent with the word2vec and GloVe-based chained weighted-sum models, where the models with greater weights on the verb and noun were most predictive of Cloze responses.

Table 4.4 displays the explained variance for the best-fitting weighted-sum models predicting Cloze z-RTs.

Table 4.4. Explained variance in Cloze z-RTs in weighted-sum models

Structural Model	**Chained/ Unchained**	**Cloze z-RTs: Highest Fixed R^2/Total-R^2 (%)**	**Best Weighted Sum Combination/Fragment (Adjective-Noun-Verb)**
ELP	-	5.85 [1.91, 6.37]/14.65	-
word2vec	Unchained*	8.08 [3.79, 8.82]/14.41	0.1+0.1+0.8
	Chained	7.48 [2.76, 7.75]/15.72	0.1+0.1+0.8

GloVe	Unchained*	7.91[3.14, 8.28]/16.26	0.1+0.1+0.8
	Chained	7.10 [2.55, 7.39]/15.57	0.1+0.1+0.8
BERT*		9.65 [5.24, 9.96]/16.05	The adj-noun-verb-the
		10.78 [6.30, 11.24]/16.11	**adj-noun-verb-the**
		9.78 [5.41, 10.16]/16.62	noun-verb-the
		7.31[2.72, 7.38]/15.34	verb-the
		6.29 [2.02, 6.68]/14.99	The

Note: * indicates significant (p < .05) increase in variance based on log-likelihood tests over the previous model. CI indicates the 95% confidence interval based on bootstrapped R^2 estimates from 1000 samples with replacement.

The weighting for the z-RT models that produced the best fit was 0.1 (adjective)+ 0.1 (noun) + 0.8 (verb) across both word2vec and GloVe, which placed far greater emphasis on the verb, compared to both the noun and adjective in predicting response latencies. Again, this is generally consistent with the response probabilities, and indicates that the verb was critical in predicting task performance. These results were also consistent with the truncated fragment-based results from BERT, where removing the verb led to the sharpest drop in explained variance (p = .002), again suggesting that the verb was more critical in determining Cloze z-RTs.

4.6 Discussion

The analyses of Cloze responses and latencies yielded four important findings. First, it is noteworthy that one can get pretty far in predicting response probabilities simply by examining item-level variables from the ELP. Second, an attention-based contextual DSM, BERT, significantly outperformed other distributional models (word2vec and GloVe) in accounting for Cloze task performance. Third, incorporation of the difference between response and competitor activations and the mean neighbor activations improved the predictive power for all models. Therefore, a delta-based thresholding process combined with overall mean neighbor activations likely influenced Cloze responses, such that once a response was sufficiently more activated than the competitor, it was more likely to be selected as the final response, and greater neighboring activations facilitated response production. Finally, the analyses of the contribution of different content words showed that the verb primarily influenced the likelihood of a given response as well as the response latencies to produce a response. Each of these findings is now discussed in detail below.

An interesting finding from these analyses is that simple ELP-based variables strongly predicted Cloze responses. Indeed, there was relatively little added variance above and beyond these variables for some of the simple models based on the non-contextual DSMs. This is an important observation because it clearly indicates a benchmark observation that is rarely tested in model evaluation, i.e., item-level relationships between retrieval cues and responses influence response production across a variety of tasks, and it is therefore important to control for these variables and lexical biases when evaluating different semantic models.

The analyses across different structural DSMs suggested that an attention-based contextual DSM, BERT, showed the best performance in predicting Cloze responses and latencies, compared to word2vec and GloVe. Although this is consistent with prior work in the natural language processing literature, where BERT has been shown to outperform several other DSMs in different semantic tasks (Devlin et al., 2019), no work has examined the extent to which BERT can account for either behavioral response probabilities or response latencies in the Cloze task. The present results suggest that BERT's modeling framework is powerful enough to not only provide sensible completions to sentences, but also predict the time course of human-generated completions. An important question regarding these findings is regarding the nature of the representation contributing to these effects. Specifically, the BERTforMaskedLM model (used in this work) is explicitly trained to perform the task of predicting missing words in a sentence based on sentences derived from very large text corpora. Therefore, one may not be surprised that BERT successfully outperforms other models in the Cloze task. However, it is important to reiterate that word2vec is *also* a predictive neural network like BERT, and derives its semantic representations by predicting words within a prespecified context window. The mechanism unique to BERT is the weighting that it assigns to different words during this prediction process, which ultimately leads to more powerful semantic representations of the words embedded within different sentence contexts. Therefore, these findings provide cognitive support to the model architecture of BERT, and specifically the weighting-based "attention" mechanism that contributes towards BERT's predictions.

The analyses of truncated sentences provided to BERT in the present study also show that BERT is sensitive to specific parts of the Cloze fragment and uses this information to generate prediction scores, and these predictions become weaker as the information supplied to BERT is systematically reduced. Of course, the scale at which BERT is trained (i.e., the text corpora) as well as the specific finetuning of parameters that enable BERT to make these predictions are important factors that likely also contribute to this performance (see Kumar, 2020 for a discussion). There is also recent work that shows that although BERT may be significantly better than previous models in predicting responses in the Cloze task, it is has considerable difficulty in predicting human performance in generating sensible inferences, responding to negation, and other language processing behavior that comes very naturally to most individuals (Ettinger, 2020; Niven & Kao, 2019). Indeed, even within the current dataset (CLOZE-338), as shown in Table 4.5, BERT did not always correctly predict the modal response and often generated odd predictions, even when the modal response was produced by most participants in CLOZE-338.

Table 4.5. Examples of incorrect BERT predictions in CLOZE-338

Cloze Fragment	**Modal Response (Probability)**	**BERT prediction**
The helpful librarian ordered the	book (.97)	search
The little mouse ate the	cheese (.92)	cat
The dirty dog buried the	bone (.89)	body
The oily dressing ruined the	salad (.87)	effect
The pastry chef decorated the	cake (.78)	room

Indeed, BERT correctly predicted the *modal* Cloze response only 27% of the times within the current dataset and generated reasonable predictions only 71% of the times[10], clearly indicating that one of the best language models has a considerable way to go to capture human Cloze task performance. Qualitative analyses showed that incorrect BERT predictions were generally driven by stereotypical sub-fragment completions (e.g., buried the body) or frequently co-occurring nouns (e.g., cat-mouse) that overwhelmed other parts of the fragment. Of course, there may also be some limitations with respect to the reliability of the CLOZE-388, which is likely to place a ceiling on explained variance.

Another important finding from the present study was that the verb strongly contributed to the response activation and selection process in the Cloze task, whereas the contribution of the noun was slightly higher in response probabilities, and the adjective contributed minimally to task performance. This relative weighting of different content words is similar to the positional weighting that BERT implements within a neural network predicting words going left to right and right to left, and provides a computational account of how individuals may process incoming words and assign differential emphasis on these words based on the task demands. For example, consider the fragment “the young landscaper mowed the” - although the adjective (*young*) and

[10] Judgments of reasonability were obtained from 2 independent raters who scored whether a human would complete the fragment with the BERT prediction as a 0 or 1. Rater judgments were moderately correlated, r = .57 and averaged to .70 and .73 respectively.

noun (*landscaper*) may help in activating the relevant semantic space from which a specific response may be selected (e.g., words related to gardening), it is the verb that determines the suitability of the response (e.g., *lawn*) to the syntactic and semantic structure of the fragment, which in turn affects the response latencies. It is important to note here that both word2vec and GloVe also emphasized the importance of the verb. Therefore, future work should focus on more carefully exploring how different parts of a sentence fragment influence the time course of Cloze response production.

The present results also provided some insight into the influence of neighboring activations influence Cloze response production and latencies. The finding that delta between the response and competitor predicts Cloze task performance is consistent with findings from the free association task (Chapter 3) and clearly demonstrates that in language-production tasks there is competition amongst different activated representations. A process that captures this competition is critical in accounting for the likelihood of a specific response being selected as well as the time taken to produce that response. Delta also strongly predicted an important behavioral pattern within the Cloze task - Cloze responses were produced faster in more constraining fragments. Additionally, mean neighbor activations also contributed to this pattern, such that activation of related words within the semantic space actually facilitated response selection and production. This is in contrast to the effect of mean neighbor activations on free association responses in Chapter 3, where an inhibitory effect of highly activated neighbors was observed. Indeed, it is possible that when linguistic context is constraining enough (in the form of a five-word fragment in the Cloze task), the activation of other words in the semantic space helps in selecting the appropriate response by spreading more activation to the more likely completions. On the other hand, in the free association task where the context (in the form of a one-word cue) is not sufficiently constraining, high activation of several words in fact interferes with response selection. This finding again highlights how different tasks may tap into different operations and underscores the need to develop task-specific computational models when accounting for behavior, although more work is needed to fully understand the influence of neighboring activations on response production in different language tasks. Collectively, these analyses provide a computational method to capture activations within semantic space via algorithmic models based on DSMs and ultimately account for behavioral patterns in the Cloze task.

In sum, the present study investigated the extent to which representations derived from distributional semantic models, when combined with appropriate algorithmic assumptions account for responses and latencies in the Cloze task. These findings highlight how syntactic and semantic information within fragments is critical to the process of selecting a Cloze response, and how a *contextual* attention-based model of semantic memory provides a better account of Cloze task performance, compared to non-contextual distributional models. Furthermore, accounting for response competition improved model fit, suggesting that Cloze response selection also involves attending to competing words in the semantic neighborhood. In this vein, this chapter provided a novel computational account of combining distributional models with algorithmic models to account for behavior in another standard language production task.

Chapter 5:
General Discussion

This dissertation delineated a novel computational approach to model retrieval processes in two familiarity-driven tasks (relatedness and similarity judgments) and two language production tasks (free association and the Cloze task), by combining different distributional models of semantic memory with algorithmic and processing principles to predict responses and response latencies within each task. This chapter discusses the primary findings from each chapter, and how these findings complement the prior literature and further inform our understanding of how semantic information guides behavior.

5.1 Predicting Relatedness and Similarity Judgments

Chapter 2 explored how semantic representations derived from two distributional models (word2vec and GloVe) when combined with a process-level model (proposed by Rotaru et al., 2018) and different algorithmic models accounted for relatedness and similarity judgments in the MEN/SimLex-999 datasets. The analyses suggested that the Rotaru et al. process model significantly outperformed all other algorithmic models in capturing both relatedness and similarity judgments, explaining variance over and above ELP variables in this task. Importantly, the Rotaru et al. model was based on the spreading activation mechanism and captured overall activation levels across time within a semantic space, which effectively accounted for performance in these tasks likely because they are driven via familiarity-based processes. In addition, these findings suggest that algorithmic models that focus on a selection process amongst competing activations within the semantic space may not be particularly suited towards tasks that are driven by overall activations. Indeed, as recent work on modeling response latencies in relatedness judgments (Kraemer et al., 2021) suggests, judgements of relatedness (and similarity) may be driven by overall semantic relatedness levels and decision processes that accumulate evidence for one decision (related) over another (unrelated) across time. However, such decisions may not require explicit modeling of other activated competitors given that the task does not require the explicit selection of a particular word, as in other language production tasks. Future work should examine how semantic representational accounts integrate with such decision-based processes, and compare to the Rotaru et al. model in accounting for relatedness judgments. A joint examination of relatedness/similarity judgments as well as response latency data is critical to this enterprise. Overall, however, the first chapter demonstrated how distributional semantic models, when combined with appropriate process-level assumptions based on the spreading activation mechanism, successfully accounted for behavioral patterns in relatedness/similarity judgments. Given that relatedness/similarity judgments may reflect more familiarity-driven processing, Chapters 3 and 4 examined whether alternative algorithmic models that take into account competing activations would better account for performance in more attention-demanding language production tasks.

5.2 Predicting Free Association Responses and Latencies

Chapter 3 focused on free association responses and latencies from the Small World of Words database. The analyses suggested that an algorithmic model based on multiplying semantic similarity by frequency and comparing competing neighbor activations significantly outperformed the Rotaru et al. process-driven model in accounting for free association responses and latencies. These results suggest that familiarity-driven tasks tap into different operations that may depend on overall activation levels within a semantic space, compared to production-based tasks such as free association that instead require the selection of a single response amongst different competitors. Indeed, the analyses of delta-based models in Chapter 3 suggested that there may be a threshold above which sufficiently activated responses are more likely to be produced during free association. Importantly, these delta-based effects were not limited to response likelihoods, but also explained response latencies, such that responses were produced faster if they were sufficiently more activated than a competing word and the overall activation of other neighbors within the semantic space was low.

Chapter 3 also provided a novel computational account for how secondary responses may be produced in a continued free association task. The empirical comparisons between a chained and unchained model showed that the chained model that emphasized additional activation from the primary response best predicted secondary response likelihoods and latencies, compared to a model that emphasized additional cue-based activations. Furthermore, incorporating delta and neighbors-based variables further improved the fit of the chained model. Collectively, these findings indicate that secondary response production is dependent on primary responses as well as competing activations of neighbors within that semantic space, providing further support for the hypothesis that language production tasks require the explicit modeling of response selection amongst competitors. Importantly, although previous work has demonstrated the presence of chaining in the SWOW database (De Deyne et al., 2019), the current work represents the first account of how such chaining actually occurs within the free association task, and how primary response activations influence secondary responses and latencies.

Overall, Chapter 3 demonstrated how the mechanisms underlying free association involve complex interactions between different sources of information. Indeed, the analyses suggest that free associations draw on information such as word concreteness and valence in addition to activations driven by semantic similarity and frequency, and that these varied sources of information combine to produce the final response during free association. The present work provides the first exploration of the time course of response production during free association, based on distributional semantic models trained on text corpora. Although the analyses suggest reliable effects of semantic similarity, and competitor as well as neighbor activations derived from these text corpora, it is important to highlight that there is still considerable variance to be explained within these data, especially within RTs. Given the multimodal nature of free associations (De Deyne et al., 2021), future work should examine how a multimodal model (i.e., a model that takes into account sensorimotor information that humans are exposed to in the natural environment) may account for free association responses and latencies, when integrated with appropriate process-level or algorithmic models of response production.

5.3 Predicting Cloze Responses and Latencies

Chapter 4 explored how different algorithmic models accounted for performance in the Cloze task, another common language production task. The analyses indicated that the contextualized BERT model, which is based on position and context-based "attentional" weighing mechanisms, significantly outperformed the word2vec and GloVe models in accounting for Cloze responses and z-RTs. Furthermore, the parametric analyses indicated that different components of the Cloze fragment (i.e., adjective, noun, and verb) differentially contributed to response production, confirming the hypothesis that the process of producing a response in the Cloze task involves attending to the critical content words in a syntactically constrained manner. Finally, similar to free associations in Chapter 3, incorporating competitor and neighbor activations via a delta-neighbors model significantly improved the predictive power of all models and also successfully accounted for critical z-RT patterns in the Staub et al. dataset, providing further support for the hypothesis that language production tasks involve the selection of responses amongst different activated words, and the extent to which a response is more or less activated than its competitors determines its response likelihood as well as response latencies.

Despite BERT significantly outperforming all other models in predicting Cloze task performance, it is important to highlight here that the explicit predictions within BERT were still far off from the human baseline and the variance explained in Cloze response latencies was overall low. This suggests that even the seemingly effortless task of predicting the end of a sentence is driven by complex syntactic and semantic interactions, in conjunction with general knowledge-based inferences, and state-of-the-art language models have difficulty accounting for behavioral data without explicit access to such world knowledge. For example, within the Cloze task, a human would almost never complete the fragment "the little mouse ate the" with the word "cat" (which was BERT's prediction), because individuals *know* that it is nearly impossible for a mouse to eat a cat due to their relative sizes and the prey-predator relationship these two animals share. Moreover, although BERT uses position and linguistic context to assign weights to different content words (i.e., "attention"), it remains unclear whether and how this notion of weighting truly captures the cognitive construct of attention. Overall, there is a need to further understand how individuals process incoming information to make inferences on the fly, as well as how individuals apply knowledge schemas to language-based tasks. Current work in natural language processing is attempting to incorporate knowledge graphs and cause-and-effect relations gathered via crowdsourced databases such as ConceptNet (Speer et al., 2017) and COMET (Bosselut et al., 2019) within distributional models to fully understand the strengths and limitations of current language models in accounting for behavioral benchmarks[11], which would ultimately inform some of these research questions.

5.4 Task Specificity in Model Evaluation

One important theme that emerges from the present work is that the specific *task* on which different models are evaluated is critical. For example, the Rotaru et al. model was specifically developed to account for dynamics of retrieval in familiarity-based tasks. Therefore, although the present work examined how this model could explain production-based task performance, it is possible that the general Rotaru et al. framework could be extended or modified to incorporate delta-based competitor activations and ultimately account for processes that may be relevant to tasks such as free association and sentence completion. Furthermore, the present work shows how the efficacy of a model is strongly dependent on *how* it is evaluated. For example, although BERT may capture performance in the Cloze task (which, as the current work shows is still very far from the human baseline), how would it account for free association, a task that does not rely on sentential context? The current literature is rife with examples of models evaluated on very different tasks (ranging from simple tasks such as assessing word similarity to complex tasks such as reading comprehension) but these tasks may not necessarily reflect the same *cognitive* demands or principles. Indeed, as the present work highlights, different tasks demand different types of underlying mechanisms. Therefore, if a model (e.g., BERT) is *finetuned* to achieve state-of-the-art performance on a question answering database, this reflects the ability of the model to *learn* a behavior after several rounds of training (which may be important for

[11] See https://homes.cs.washington.edu/~msap/acl2020-commonsense/ for a recent workshop on commonsense reasoning within state-of-the-art language models

developing language-based technologies), as opposed to the flexibility that human processing systems possess to modify their processes based on task goals, with minimal prior training. Ultimately, modeling this *flexibility* of the cognitive system to perform virtually *any* task will be critical for the natural language processing enterprise, and understanding on how different tasks demand different cognitive processing is an important next step for the field (see Balota & Yap, 2006).

On a related note, it is important to highlight the role of item-level ELP variables in accounting for performance across different semantic tasks. In the present work, variables such as length, frequency, concreteness, and valence accounted for relatively large amounts of variance in predicting responses and response latencies across all tasks. Indeed, surprisingly, the *gains* in accounted variance from semantic variables based on DSMs were small (albeit significant) in some tasks (e.g., Cloze task) compared to others (e.g., relatedness judgments), which suggests that it is important to separate the contribution of these ELP-based variables from information derived via DSMs in accounting for performance across different tasks. As discussed, one possibility is that these ELP variables may reflect natural lexical biases that influence behavioral performance within semantic tasks. Work in machine learning and natural language processing typically does not control for lexical variables when assessing model adequacy, but the present work suggests that it is important to assess the predictive power of semantic models relative to the influence of item-level characteristics, especially because different tasks may differentially emphasize these item-level relationships. A second possibility is that the ELP variables and the DSMs contain complementary *semantic* information that may be critical in semantic tasks. Indeed, as evidenced by low correlations between cosine similarities from DSMs and these ELP variables, it is possible that DSMs trained on linguistic corpora may be particularly disadvantaged at capturing information about concreteness or valence, which may be important in tasks such as similarity judgments or free association. In this light, one may assert that DSMs do *not* perfectly represent word meaning, and ELP variables may actually represent an important component of what constitutes meaning. Finally, a third possibility is that there is shared variance between tasks used to obtain ELP information and the tasks examined in the present dissertation, which may be a contributing factor to the variance explained by these variables. Although the present work cannot fully discriminate between these possibilities, it is important to highlight that the ELP variables may simultaneously represent natural lexical biases, as well as non-linguistic meaning-related information that may be critical when accounting for performance in semantic tasks.

Collectively, this dissertation described an approach to model behavior in semantic retrieval tasks, by combining distributional semantic models that learn semantic word meaning via large-scale text corpora with appropriate algorithmic/process-based models. An important takeaway from these studies is that different tasks tap into different *processing* operations, and production-based semantic tasks (such as free association and the Cloze task) require explicit modeling of search and retrieval processes leading up to the identification of a single response that are different from familiarity-based tasks (such as relatedness/similarity judgments) that may instead depend on accrual of activation between concepts within a semantic space. Furthermore, the present work highlights that different *representational* models provide the best account of performance across the different tasks (i.e., GloVe in relatedness/similarity judgments and free association, and BERT in the Cloze task). A complete account of semantic memory must be able to accommodate these representations and processing operations within the same general framework. Although it is unlikely that there are multiple semantic memory *systems* within the brain, it is possible that meaning is indeed represented via different *patterns* of activation when embedded within strongly constraining contexts versus unconstrained contexts, as in neurally inspired models of cognition (e.g., McClelland & Rumelhart, 1981). Indeed, it may be the case that each instance of semantic retrieval is in fact a distinct pattern of neural activity (Musz & Thompson-Schill, 2015), and specific task demands control the extent to which these patterns are activated and modified. Just as different measures of word recognition bring online distinct processes (see Balota, Spieler, & Paul, 1999), different measures of semantic memory may bring online distinct processes. Clearly, there is a long way to go before distributional models can be considered perfect accounts of semantic memory representation and processing. However, as the present work demonstrated, assuming meaning is derived via statistical regularities in natural language, indexing activations as a function of semantic similarity and frequency, and accounting for relative activations within a semantic space derived via distributional models can serve as a powerful computational account for how information is learned, accessed, and produced within language tasks. Of course, it is critical to note here that despite consistent evidence for specific models explaining the most variance in the different tasks evaluated in this dissertation, overall explained variance in production tasks (Chapters 3 and 4) was low with small (albeit significant) differences across the models. This suggests that capturing the variability in language production tasks remains an important challenge for computational semantic models overall, and future work should explore how different models can accommodate this variability.

5.5 Limitations and Future Directions

Despite the promise of the present approach, it is important to acknowledge some limitations of this work as well as discuss some future directions. First, the present analyses were carried out on secondary data for relatedness/similarity judgments, free association, and Cloze tasks, and these data were not specifically collected to explore how distributional information, when combined with algorithmic models, accounts for responses and response latencies. For instance, as the current analyses show, there was considerable variability across RTs in the SWOW database, given that participants were not encouraged to respond quickly and also performed the study through different operating systems and devices. This limits the extent to which RTs can be modeled for free association using the present data, although the present work shows how algorithmic models that index activation levels and competition within the semantic space via distributional information do indeed account for critical patterns in RTs. A critical assumption here is that even under un-speeded conditions, response latencies can be informative and likely reflect when sufficient information has been accrued within a given task to make a response. In contrast, the sentence fragments used in the Cloze data were specifically controlled to reduce item variability and followed the same sentence structure (i.e., The-adjective-noun-verb-the). Indeed, the current analyses show systematic effects of this structure on performance, in that individuals attended to verbs more than nouns and adjectives while generating sentence completions, and this behavior was mirrored in the DSMs. However, it was not possible to explicitly examine the effects of different syntactic structures, given that the same noun did not occur with different verbs and vice versa, which may provide further insight into how fragment constraint modulates performance in the Cloze task. Finally, although MEN/SimLex-999 are considered benchmark datasets in machine learning, these datasets reflect very different types of tasks (i.e., MEN asked participants to compare two word-pairs and select the more related pair whereas SimLex-999 asked participants to produce a similarity score for a single word-pair) and also do not report response latencies. RTs within cognitive tasks provide important constraints for process-level and algorithmic models and therefore the lack of RTs for these data limits the conclusions one can draw from the present analyses. For example, relatedness decisions tend to show critical patterns in RTs (e.g., the inverted U effect, see Kenett et al., 2017 and Kumar, Balota, & Steyvers, 2019), which provide important benchmarks for different process/algorithmic models.

Given that this dissertation only compared a limited set of models (word2vec, GloVe, and BERT), it cannot speak to the predictive power of these models against several other competing models in the field (e.g., Topic models, retrieval-based models, recurrent neural networks etc.) in these particular tasks. Further, although all the models tested in this dissertation were all trained on Wikipedia corpora, they were not trained on the same *size* of corpora (e.g., word2vec and BERT were trained on ~3 billion tokens and GloVe on 6 billion tokens) and some models also had some additional training on other corpora (e.g., BERT was trained on an additional BooksCorpus), and therefore some model-based differences could be attributed to corpora-level differences. However, training size does not appear to be the only discriminating factor between the performance of word2vec and BERT, because both models were trained on a corpus of approximately 3 billion tokens. Moreover, there was considerable task specificity for the different distributional models, wherein GloVe (trained on a 6 billion corpus) did poorly on Cloze task data in comparison to BERT (trained on a smaller corpus) but relatively well on the SWOW data. Therefore, the underlying mechanisms for these different models appear to have important implications for the tasks they can effectively account for, above and beyond the corpora-level differences. However, given the overall consistency in model fits across multiple tasks, the present analyses do provide overall support for the claim that distributional models can indeed be applied to language production tasks.

Additionally, the present dissertation considered only one process-based model (based on Rotaru et al., 2018), and it is possible that other process-level models (e.g., drift-diffusion, accumulator models, etc.) may also be viable candidates for explaining performance across language production tasks. Of course, the present dissertation focused on *one* possible general framework of combining distributional models with algorithmic models that may be applicable to unconstrained semantic tasks and exploring alternative process-based models is an avenue for future work. Along similar lines, the present work focused on a *spreading activation* metaphor with concepts being represented in a vector-based "network" to operationalize the amount of activation between any two concepts. Of course, another way to think about relatedness between concepts is to consider the overlap in semantic *features*, such as in Plaut and Booth's (2000) model of semantic priming. Indeed, feature-based models of semantic memory have historically informed theories of semantic memory structure and processing (McRae, 2004) and integrating feature-based information with linguistic sources of information is a thriving area of research (Jones, Willits, & Dennis, 2014). Therefore, future work should compare different ways of indexing semantic activation between concepts to further understand how concepts are accessed and retrieved from semantic memory.

It is also important to highlight here that the present models were all deterministic, i.e., each model produced a *single* prediction for a given item, based on one underlying semantic representation. Furthermore, although this is the standard approach in this literature, using pretrained models automatically limits the finetuning of certain model parameters that may be able to capture some type of individual-level variation. Therefore, a simplifying assumption within the present work was that all individuals share the *same* semantic memory system. However, as the present work demonstrates, individuals exhibit considerable variability in responses and response latencies, and an accurate computational model of semantic memory and language production should be able to account for this variability. Indeed, future work could explore individual-level generative Bayesian representational models (e.g., as in Zemla, Kenett, Jun, & Austerweil, 2016) as well as stochastic noise-based algorithmic/process models, to develop more accurate models of retrieval from semantic memory.

5.6 Conclusions

Retrieval from semantic memory is fundamental to all cognition, and therefore it is crucial to understand how information is searched, retrieved, and ultimately used to produce a response in a cognitive task. The present work introduced a quantitative framework based on combining distributional semantic representations with different algorithmic models and one process-based model to account for performance in two familiarity-based tasks (relatedness and similarity judgments) and two language production tasks (free association and sentence completion). The analyses showed that production tasks involve complex interactions between activations of different competing words within a semantic space to produce a single response, whereas familiarity-driven tasks are influenced by overall activation of concepts within a semantic space. The present studies build upon the previous literature by identifying specific mechanisms of activation and competition by which a specific response is selected and produced in response to task-dependent cues, and also suggest important differences across different semantic tasks and their underlying processing mechanisms. Ultimately, the framework from this dissertation could be extended to explore how individuals perform more open-ended production-based semantic tasks such as question answering and lexical retrieval, therefore introducing a quantitative approach to understanding the dynamics of retrieval processes across different semantic tasks.

References

Abbott, J. T., Austerweil, J. L., & Griffiths, T. L. (2015, July). Random walks on semantic networks can resemble optimal foraging. In *Neural Information Processing Systems Conference. 22*(3). 558. American Psychological Association.

Alammar, J. (2018). The Illustrated Transformer [Blog post]. Retrieved from https://jalammar.github.io/illustrated-transformer/

Alm, C. O., Roth, D., & Sproat, R. (2005, October). Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pp. 579–586, Vancouver, October 2005.

Anderson, J. R. (1996). ACT: A simple theory of complex cognition. *American Psychologist*, *51*(4), 355.

Anderson, J. R., & Bower, G. H. (1974). A propositional theory of recognition memory. *Memory & Cognition*, *2*(3), 406-412.

Aschenbrenner, A. J., Balota, D. A., Gordon, B. A., Ratcliff, R., & Morris, J. C. (2016). A diffusion model analysis of episodic recognition in preclinical individuals with a family history for Alzheimer's disease: The adult children study. *Neuropsychology*, *30*(2), 225.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance, 10*(3), 340.

Balota, D. A., & Yap, M. J. (2006). Attentional control and the flexible lexical processor: Explorations of the magic moment of word recognition. *From Inkmarks to Ideas: Current Issues in Lexical Processing*, *229*.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., ... & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*(3), 445-459.

Balota, D. A., & Lorch, R. F. (1986). Depth of automatic spreading activation: Mediated priming effects in pronunciation but not in lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*(3), 336.

Balota, D. A., Paul, S. T., & Spieler, D. H. (1999). Attentional control of lexical processing pathways during word recognition and reading. *Language processing*, pp. 15-57.

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

Barton, K. (2020). MuMIn. R package. version 1.43.17.

Bhatia, S., & Richie, R. (under review). Transformer networks of human concept knowledge. PsyArXiv preprint.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Agarwal, S. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*. Retrieved from https://arxiv.org/pdf/2005.14165.pdf.

Bruni, E., Tran, N. K., & Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, *49*, 1-47.

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, *39*(3), 510-526.

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

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *8*(2), 240-247.

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*,160-167. ACM.

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods*, *51*(3), 987-1006.

De Deyne, S., Perfors, A., & Navarro, D. J. (2016, December). Predicting human similarity judgments with distributional models: The value of word associations. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 1861-1870).

De Deyne, S., Navarro, D. J., Collell, G., & Perfors, A. (2021). Visual and Affective Multimodal Models of Word Meaning in Language and Mind. *Cognitive Science*, *45*(1), e12922.

De Deyne, S., & Storms, G. (2008). Word associations: Network and semantic properties. *Behavior Research Methods*, *40*(1), 213-231.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*(3), 283.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Diedenhofen, B., & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PloS one*, *10*(4), e0121945.

Ettinger, A. (2020). What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, *8,* 34-48.

Faust, M. E., Balota, D. A., Spieler, D. H., & Ferraro, F. R. (1999). Individual differences in information-processing rate and amount: implications for group differences in response latency. *Psychological Bulletin*, *125*(6), 777.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. In Philological Society (Great Britain) (Ed.), *Studies in Linguistic Analysis*. Oxford: Blackwell.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review, 114*(2), 211.

Gruenenfelder, T. M., Recchia, G., Rubin, T., & Jones, M. N. (2016). Graph-theoretic properties of networks based on word association norms: implications for models of lexical semantic memory. *Cognitive Science*, *40*(6), 1460-1495.

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science*, *14*(6), 1006-1033.

Harris, Z. (1970). Distributional structure. In *Papers in Structural and Transformational Linguistics* (pp. 775-794). Dordrecht, Holland: D. Reidel Publishing Company.

Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4), 665-695.

Hutchison, K. A., Balota, D. A., Cortese, M. J., & Watson, J. M. (2008). Predicting semantic priming at the item level. *Quarterly Journal of Experimental Psychology, 61*(7), 1036-1066.

Jones, M. N., Gruenenfelder, T. M., & Recchia, G. (2018). In defense of spatial models of semantic representation. *New Ideas in Psychology, 50,* 54-60.

Jones, M. N., Hills, T. T., & Todd, P. M. (2015). Hidden processes in structural representations: A reply to Abbott, Austerweil, and Griffiths (2015). *Psychological Review*, *122*(3), 570–574. doi:10.1037/a0039248

Jones, M. N., Willits, J., Dennis, S., & Jones, M. (2015). Models of semantic memory. *Oxford Handbook of Mathematical and Computational Psychology*, 232-254.

Kenett, Y. N., Anaki, D., & Faust, M. (2014). Investigating the structure of semantic networks in low and high creative persons. *Frontiers in Human Neuroscience, 8,* 407.

Kenett, Y. N., Kenett, D. Y., Ben-Jacob, E., & Faust, M. (2011). Global and local features of semantic networks: Evidence from the Hebrew mental lexicon. *PloS one*, *6*(8), e23912.

Kenett, Y. N., Levi, E., Anaki, D., & Faust, M. (2017). The semantic distance task: Quantifying semantic distance with semantic network path length. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(9), 1470.

Kounios, J., & Holcomb, P. J. (1994). Concreteness effects in semantic processing: ERP evidence supporting dual-coding theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 804.

Kraemer, P. M., Wulff, D. U., & Gluth, S. (2021). A sequential sampling account of semantic relatedness decisions. https://doi.org/10.31234/osf.io/ksa2g

Kumar, A.A. (2021). Semantic Memory: A Review of Methods, Models, and Current Challenges. *Psychonomic Bulletin & Review*, *28*, pp. 40-80.

Kumar, A. A., & Balota, D. A. (2020). Attempted prime retrieval is a double-edged sword: Facilitation and disruption in repeated lexical retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(8), 1505–1532. https://doi.org/10.1037/xlm0000827

Kumar, A. A., Balota, D. A., & Steyvers, M. (2020). Distant connectivity and multiple-step priming in large-scale semantic networks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46(*12), 2261 -2276. https://doi.org/10.1037/xlm0000793

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.

Lazaridou, A., Pham, N. T., & Baroni, M. (2015). Combining language and vision with a multimodal skip-gram model. *arXiv preprint arXiv:1501.02598*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*. Retrieved from https://arxiv.org/abs/1907.11692.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. Courier Corporation.

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57-78.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review*, *88*(5), 375.

McRae, K. (2004). Semantic memory: Some insights from feature-based connectionist attractor networks. *The Psychology of Learning and Motivation: Advances in Research and Theory*, *45*, 41-86.

McRae, K., Khalkhali, S., & Hare, M. (2012). Semantic and associative relations: Examining a tenuous dichotomy. In V. F. Reyna, S. B. Chapman, M. R. Dougherty, & J. Confrey (Eds.), *The Adolescent Brain: Learning, Reasoning, and Decision Making* (pp. 39-66). Washington, DC: APA.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Musz, E., & Thompson-Schill, S. L. (2015). Semantic variability predicts neural variability of object concepts. *Neuropsychologia*, *76*, 41-51.

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General*, *106*(3), 226.

Neely, J. H. (2012). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In *Basic processes in reading* (pp. 272-344). Routledge.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402-407.

Nematzadeh, A., Meylan, S. C., & Griffiths, T. L. (2017). Evaluating Vector-Space Models of Word Representation, or, The Unreasonable Effectiveness of Counting Words Near Other Words. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*

Niven, T., & Kao, H. Y. (2019). Probing neural network comprehension of natural language arguments. *arXiv preprint arXiv:1907.07355*. Retrieved from https://arxiv.org/pdf/1907.07355.pdf.

Patel, A., Sands, A., Callison-Burch, C., & Apidianaki, M. (2018). Magnitude: A fast, efficient universal vector embedding utility package. *arXiv preprint arXiv:1810.11190*. Retrieved from https://arxiv.org/abs/1810.11190.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (EMNLP) (pp. 1532-1543).

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

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.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8). Retrieved from https://www.techbooky.com/wp-content/uploads/2019/02/Better-Language-Models-and-Their-Implications.pdf.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873-922.

Rayner, K., Ashby, J., Pollatsek, A., & Reichle, E. D. (2004). The effects of frequency and predictability on eye fixations in reading: implications for the EZ Reader model. *Journal of Experimental Psychology: Human Perception and Performance, 30*(4), 720.

Rayner, K., & Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin & Review*, *3*(4), 504-509.

Richie, R., Aka, A., & Bhatia, S. (in prep). Free association in bidirectional memory networks.

Rotaru, A. S., Vigliocco, G., & Frank, S. L. (2018). Modeling the Structure and Dynamics of Semantic Processing. *Cognitive Science*, *42*(8), 2890-2917.

Sheridan, H., & Reingold, E. M. (2012). The time course of predictability effects in reading: Evidence from a survival analysis of fixation durations. *Visual Cognition*, *20*(7), 733-745.

Singh, M., Richie, R., & Bhatia, S. (2020). Representing and Predicting Everyday Behavior.

Smith, N., & Levy, R. (2011). Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 33, No. 33).

Staub, A., Grant, M., Astheimer, L., & Cohen, A. (2015). The influence of cloze probability and item constraint on cloze task response time. *Journal of Memory and Language*, *82*, 1-17.

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.

Taylor, W. L. (1953). “Cloze procedure”: A new tool for measuring readability. *Journalism Quarterly*, *30*(4), 415-433.

Thawani, A., Srivastava, B., & Singh, A. (2019, June). SWOW-8500: Word association task for intrinsic evaluation of word embeddings. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP* (pp. 43-51).

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, 141-188.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychological Review*, *108*(3), 550.

Westerink, J. H., Van Den Broek, E. L., Schut, M. H., Van Herk, J., & Tuinenbreijer, K. (2008). Computing emotion awareness through galvanic skin response and facial electromyography. In *Probing Experience* (pp. 149-162). Springer, Dordrecht.

Yamada, I., Asai, A., Shindo, H., Takeda, H., & Takefuji, Y. (2018). Wikipedia2Vec: an optimized tool for learning embeddings of words and entities from Wikipedia. *arXiv preprint arXiv:1812.06280*. Retrieved from https://arxiv.org/abs/1812.06280

Zemla, J.C., Kennett, Y.N., Jun, K-S., & Austerweil, J.L. (2016). U-INVITE: Estimating individual semantic networks from fluency data. In *Proceedings of the 38th Annual Meeting of the Cognitive Science Society* (pp. 1907-1912). Austin, TX: Cognitive Science Society.

Zou, W., & Bhatia, S. (2019). Modeling Judgment Errors in Naturalistic Numerical Estimation. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society* (pp. 3227-3233).