Defining distinctiveness: A computational and experimental analysis

by

Jackie Spear

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

In partial fulfillment of the requirements of the degree of

MASTER OF ARTS

Department of Psychology

University of Manitoba

Winnipeg

Abstract

Distinctiveness is a fundamental principle of human memory. However, definitions of distinctiveness have largely remained intuitive and imprecise (Hunt & Worthen, 2006). In Experiments 1 and 2, participants studied critical, distinctive words that were embedded in eight different categorized lists. At test, three different types of lures were presented: distinct related lures, categorical related lures, and unrelated lures. Vector-based representations of word meaning were derived using distributional models of semantics to fit the data. Namely, the Bound Encoding of the Language Environment (Jones & Mewhort, 2007) and Latent Semantic Analysis (Landauer & Dumais, 1997) were employed to derive word meaning from written text. These representations were coupled with an instance-based model of human memory, MINERVA 2 (Hintzman, 1988) to model recognition. The same experimental design as in Experiments 1 and 2 was used in Experiments 3 and 4, where DRM materials replaced the old and new category words, and where LSA and BEAGLE were used to derive the distinctive words, the lures related to those distinctive words, and the subsequent unrelated lures. Lastly, Experiment 5 aimed to formulate a priori predictions for recognition when words were sampled at random.

Keywords: computational modelling, semantic distinctiveness, natural language processing, semantic space models, computational linguistics, semantic memory, vector space models, word recognition

Acknowledgements

I would like to thank my advisor Dr. Randy Jamieson for his mentorship, encouragement and support. I would also like to thank my thesis committee members Dr. Murray Singer and Dr. Jila Ghomeshi for their wisdom and input. I would also like thank my supportive lab mates, Matt Cook and Bradley Smith for being sounding boards and offering suggestions, to my former lab mates, Dr. Chrissy Chubala and Dr. Evan Curtis for their guidance and reassurance, and to lab assistants Anna Polyvyana and Matthew Slusky for their assistance in this project. It has been and continues to be an honor to work alongside all of you, and I am grateful for the inspiration you have provided. Finally, for your patience, unconditional support and encouragement, thank you Jared Adams. Thanks for being there and for helping keep me sane. I could have not done this without you.

Defining distinctiveness: A computational and experimental analysis

People remember unusual things particularly well: a phenomenon known as the distinctiveness principle (Surprenant & Neath, 2009). The memorial benefit for unusual events is observed in recall and recognition, and holds whether items are rendered distinctive by their surface features (e.g. a word presented in a different color than other words in a list) or deep features (e.g. a word that has an unusual meaning relative to other words in a list). Distinctiveness can even be linked to our physiology and neuroanatomical cell structure. For example, visual pop-out effects have been documented, wherein reaction times are significantly faster for visual targets that are unique in the presence of homogeneous distractors (Hubel & Wiesel 1959, 1962; Treisman, 1985; Treisman & Gelade, 1980).

The von Restorff (1933) effect is perhaps the most famous example of the phenomenon, which emerged from the Gestalt school of psychology (see also Calkins, 1894). The construction of her paradigm, taken from a figure-ground kind of perspective, can be seen to directly follow from this. In her experiment, people studied words and then were tested on recall of those words. Recall for words presented in different colors and in different sizes were recalled better than words presented in the dominant color or size (Green, 1956; Wallace 1965). Her result, named the von Restorff effect has become a standard fact in research on human memory and has supported a broad range of distinctiveness results in recall.

Distinctiveness also confers a memorial benefit in recognition memory. For example, words presented in a study list in an unusual color or that differ in meaning from other words in a study list are recognized better (see Hunt, 1995 for a review).

Although it is common wisdom that distinctiveness benefits memory performance, and there are a number of characteristic experimental manipulations that are accepted as conferring

distinctiveness, debate over and a clear understanding of distinctiveness in the context of human memory performance remains. Quite often researchers have relied on their phenomenological intuitions combined with successful previous demonstrations to define what renders a stimulus distinctive. One main supposition is that any stimulus feature, shallow or deep, that distinguishes a word from others in a list renders that word distinctive: distinction by contrast. However, there is a logical error in that line of reasoning. Distinctiveness is not a property of the stimulus, but rather a corollary of that person's interpretation and processing of the stimulus. Thus, distinctiveness is a corollary of psychological processing.

Although researchers have made a great deal of progress in measuring the influence of distinctiveness on memory, and in developing a broad and deep database on the issue, one might argue that we are still at a distance from understanding what distinctiveness is, how it arises as a psychological phenomenon, and how it comes to influence memory performance. The aim of my thesis is to examine distinctiveness within a computational framework that will allow me to understand how distinctiveness arises from language processing, and how that distinctiveness is leveraged in the act of remembering, specifically in the domain of recognition memory.

To bring the problem into tractable scope, I will apply computational theories of natural language processing to derive meanings of words. I will use those representations of word meaning to simulate recognition memory performance within an established and classical theory of human memory, and I will test the extent to which that theory predicts people's performance in word recognition experiments. My analysis is motivated to show that existing theories of semantic memory combined with existing theories of recognition memory get us closer to a clear testable and articulate formal account of how semantic distinctiveness arises, and how it

influences remembering. If successful, my thesis will contribute to the larger disciplinary effort to translate an important psychological principle into a formal theoretical expression.

The Empirical Analysis of Distinctiveness in Recognition Memory

Psychology has a long history of studying the role of distinctiveness in recognition memory. Although there are many variations in how people have studied the problem, the experiments all have a common underlying structure. First, participants are asked to study a list of words in which some small number of those words are presented in a different form. For example, in a different color, or they differ from other words in the list in their meanings. In a test phase, participants' memory is tested by recognition of the words they studied, and from words that they did not. The influence of distinctiveness on recognition performance is measured by comparing recognition of study words presented in the unusual manner versus recognition of the study words presented in the typical manner. When people recognize the unusually presented words better than the typically presented words, a distinctiveness effect is observed.

The production effect is one example of a distinctiveness experiment. In a production protocol, people speak some of the words presented in the study list, but not all. At test, they show better recognition for words that they spoke than words they did not. Assuming that speaking a word that renders it distinctive in memory, the difference reveals a benefit of distinctiveness conferred by production. The result was first presented in 2010 by McLeod et al. and has been replicated a great number of times since then (Ozubko, Major, & MacLeod, 2014; Jamieson, Mewhort, & Hockley, 2016; Jamieson & Spear, 2014).

The generation effect is a second example in which scientists have examined the problem of distinctiveness. In a generation protocol people are presented with stems and completions and then are tested for their recognition for all of the completions. Critically, people are provided

with half of the completions at study (e.g. hot-cold) but must generate the completions for the others (e.g. light-d_ _ _). Assuming that generating a word renders it distinctive in memory, better recognition of generated completions over given completions reveals a benefit of distinctiveness conferred by generation. The generation effect was first documented by Slamecka and Graf (1978), and has been replicated repeatedly since then (Kornell & Terrace, 2007; McDaniel, Waddill, & Einstein, 1988).

The enactment effect is a third example of how scientists have examined distinctiveness in recognition memory. In an enactment protocol people are asked to use some objects presented in a study phase (e.g. pick up the pen). At test, people recognize the objects they used better than the objects they did not, to the extent that using an object renders memory for it distinctive. The enactment effect renders yet another example of the benefit that distinctiveness confers to memory performance (Engelkamp, 1995, 1998).

A fourth way that scientists have studied distinctiveness is by the semantic distinctiveness effect. In contrast to the production, generation and enactment effects, that examine distinctiveness by study manipulations, work on the semantic distinctiveness effect examines the benefit for a word that differs in meaning relative to other words in the same list (e.g. HAMMER presented in a list of fruits). For example, Singer, Fazaluddin, and Andrew (2011) presented people with lists of words belonging to dominant categories, but that included words drawn from outside of those categories. They also presented some of those words twice and others once. Even though they observed a standard recognition benefit for repeated words, words that did not belong to the studied categories were recognized very well after only a single presentation. Based on the results obtained by Singer et al. (2011), they concluded that semantic distinctiveness confers a strong and stable advantage in word recognition. Similar results and

conclusions have been demonstrated in other labs and by other researchers (Dewhurts, 2001; Johns, et al., 2012; Jones, Johns, & Recchia, 2012).

Taken together, the production, generation, enactment, and semantic distinctiveness effects provide strong, convincing, and converging evidence that distinctiveness, whether defined by deep semantic processing manipulations or simple study manipulations, lead to benefits in word recognition. Nevertheless, the experimental manipulations of distinctiveness remain grounded in a phenomenologically grounded understanding of how distinctiveness is defined and how it comes to confer the benefit in recognition memory performance. Although the phenomenological and norms-based approaches to defining distinctiveness have been productive and served as a sound basis for predicting distinctiveness effects in word recognition, ideally the discipline can develop a formal and competent explanation of how word meaning arises independent of a phenomenological judgement to predict when a word will be distinctive, and consequently, when it will be recognized well.

The Geometry of Meaning

Computational theories of language give us a good platform from which to examine distinctiveness. The extent to which we can leverage these theories will give us a catalyst to launch an empirical and computational account of semantic distinctiveness in human memory.

Although the computational approach only gained traction in psychology in the mid 20th century, deeming it a relatively new field, there is already a rich history of success and a good variety of approaches in cognitive science and psychology. As early as the 1950's, computational models were developed to solve reasoning problems by Newell and Simon (Simon, 1991), and in 1958, the first artificial neural network was developed (Rosenblatt, 1958). Since then, connectionist models have been developed to model language acquisition (Pinker & Prince,

1988; Rumelhart & McClelland, 1986b, 1987), Bayesian decision theory has been used to model sensorimotor control (Körding & Wolpert, 2006), and models of distributed semantics coupled with machine learning approaches have even been used to detect mental illness from language (Cook, 2018). That being said, one might question the benefits of using computational methods. To answer the criticism, Lewandowksy (1993) outlined several benefits. It forces a complete theory because everything must be specified in order to compute empirical corollaries. Models can manage difficult to comprehend interactions and lead to the discovery or development of new research directions. Most importantly, models are especially beneficial when we wish to solve complex problems that challenge the limits of human reasoning.

Scientists have worked for a long time to represent knowledge in machines. In psychology that effort started with Osgood's work with the semantic differential. In Osgood's 1957 framework, meaning is defined along principal axes (e.g. good/bad), and a word's meaning is defined by people's judgements of each word along each of those principal axes. Once each word is represented as a point in a hyper-dimensional space defined by those principal axes, similarity in meaning can be computed by the distance between the points. Osgood's initial work established the fundamentals for how a quantitative representation of meaning would be developed into the modern era.

Following on Osgood's efforts, researchers in the 1960's and 1970's represented the meaning of words in hierarchical and propositional fashions. Most famously, Collins and Quillian (1969) represented the relationships between words through branching structures that included words connected by their conceptual relationships. Those models were developed, as were Osgood's, based on the introspective judgements and assessments of experimental participants. Those methods helped the field to understand the problems of family resemblance

structure and conceptual representation, which continue to have a strong influence today (Tversky, 1977).

Following along the same tradition, researchers in the 1980's took a decidedly empirical approach to quantifying the meaning of words based on a new set of principal axes. To accomplish the goal, large numbers of university undergraduates were given the task of rating words for concreteness, imageability, meaningfulness, emotionality, pleasantness, and so on. Those ratings were catalogued and published, and an era of psycholinguistics followed to examine the relationship between language behavior and memory performance as a function of those measurements (e.g. the Toronto Word Pool and the MRC Psycholinguistic Database; Friendly, et al., 1982; Coltheart, 1981; Wilson, 1988; Gardiner & Java, 1990).

Given the availability of computers, text databases, and information theory, the field leapt forward in the 1990's. At the beginning of the decade, psychologists began to publish computational theories of natural language processing, collectively called distributional theories of semantics (e.g. Hyperspace Analogue to Language and Latent Semantic Analysis; Burgess, 1998; Lund & Burgess, 1996; Landauer & Dumais, 1997). Although those theories differ from one another in many ways, they all share a common approach and aim: to derive numeric representation of word meaning based on the statistical occurrence of word use in large bodies of text.

The theory of Latent Semantic Analysis (LSA) was the first major success. LSA, developed by Landauer and Dumais in 1997 extracts word meaning from a text database via a several step process. First, a word by document matrix is constructed, in which rows represent words and documents are represented by columns, where each cell in the matrix records the number of times each word occurs in each document. After the word by document matrix is

constructed, the word counts are transformed by one of several possible functions. Then, the matrix is decomposed by the theorem of singular value decomposition, akin to principal components analysis. Finally, a word by document matrix of reduced dimensionality is reconstructed using only the *n* largest eigenvectors, where *n* is determined by the proportion of variance that each eigenvector explains. That is, the singular values that capture the greatest amount of variance are kept from the original matrix and are used to represent each word in the corpus (Landauer & Dumais, 1997). By doing so, this method makes use of indirect inferential information to represent word meaning. Once that matrix is constructed, each row provides an *n* dimensional semantic representation of a word in the corpus, where *n* is usually set to 300. It just so happens that with this method LSA performs the best when *n* is set to 300, thus this number has become convention. Word similarity can be computed as the cosine between those rows. Words with similar meanings have a higher cosine value, words with some semantic relationship have a lower cosine value, and words with no semantic relationship are orthogonal, or have a cosine of zero. Despite the simplicity of the theory, it has been surprisingly successful at capturing people's rate of language acquisition (Landauer & Dumais, 1997), free association behavior (Steyvers, et al., 2005), priming behavior (Günther, Dudschig, & Kaup, 2016) and categorization judgements (Laham, 1997).

The Bound Encoding of the Language Environment (BEAGLE; Jones & Mewhort, 2007) is another theory of semantic memory theory grounded in Murdock's 1982 TODAM theory of distributed memory. Though LSA proved to be successful in a variety of ways, the theory suffers from the *bag of words problem*. The bag of words problem refers to the issue that a word's meaning is defined by its context *and* temporal position in that context relevant to other words. LSA cannot account for word order. As such, this was one of the main motivations for the

development of BEAGLE. BEAGLE is a model of lexical semantics. Broadly, during a simulation, it works by "reading" a text corpus and, en route, it encodes a memory vector that represents the meaning of each word in that corpus. Mechanistically, the model is expressed in algebra.

At the start of a simulation, each of the i unique words in the corpus is represented by a random environmental vector, e_i. Each environment vector has dimensionality n and each element in an environment vector obtains a value randomly sampled from a normal distribution with a mean of zero and variance of $1/n$. Most often, dimensionality is set to values between n = 1,024 and n = 2,000. In the simulations that follow, dimensionality will be set to 2,000. Environment vectors maintain stability over a simulation and are meant to serve as unique identifiers for each of the words in the corpus (i.e., both the word's orthographic and phonological identity).

Next, the model "reads" (or processes) the corpus one sentence at a time to build a semantic memory vector for each word. The memory vector for each word, m_i, is composed of two kinds of information: context information and order information. Context information is computed by summing the environmental vectors for all other words in the same sentence (i.e., excluding the word of interest). For example, after reading the sentence "A dog bit the mailman", the memory vector for *dog* is updated as $m_{dog} = m_{dog} + e_{bit} + e_{mailman}$, the memory vector for *bit* is updated as $m_{bit} = m_{bit} + e_{dog} + e_{mailman}$, and the memory vector for mailman is updated as $m_{mailman} = m_{mailman} + e_{dog} + e_{bit}$.

As should be apparent, summing the environment vectors in this way causes the memory vectors for all words in the same sentence to grow more similar to one another. Less obviously, the method also encodes higher-order associations between words. For example, even if *mailman*

and *postage* do not co-occur in the same sentence in the corpus, they will be similar to one another by virtue of having common words summed into their representations (e.g., *letter*).

Order information is accounted for by encoding information about which words follow one another in a sentence and updating the memory vector with that information. In particular, first-order association between words (i.e., immediately adjacent words) is encoded using noncommutative circular convolution; hereinafter denoted as circular convolution.

Circular convolution is a vector operation that binds two vectors, **x** and **y**, to produce an associative vector, **z**,

$$z = \sum_{j=1}^{n-1} x_{jmodn} \times y_{(i-j)modn} \quad \{\text{for } i = 0 \text{ to } n - 1\}$$

where, n is the dimensionality of **x** and **y**.

Mechanistically, a strength and convenient property of circular convolution is that it produces a vector, **z**, that is the same dimensionality as the inputs, **x** and **y**, thereby allowing the association between **x** and **y** to be summed into a single vector representation.

Of course, there exists higher-order sequential information in a sentence (e.g., sequences of three, four, or more words). Thus, to represent second-, third-, and higher-order order information, BEAGLE applies the operation recursively to update a word's order information,

$$o_i = \sum_{j=1}^{p\lambda-(p^2-p)-1} bind_{ij}$$

where, o_i is the order information for word i, p is the position of word i in the sentence, and $bind_{ij}$ is the j^{th} convolution for the word being coded.

To illustrate the operation, the order information for the word dog, o_{dog}, in the sentence, "a dog bit the mailman," is encoded as a sum of the following,

$bind_{dog,1} = e_a \circledast \Phi$ } Bigrams

$bind_{dog,2} = \Phi \circledast e_{bit}$

$bind_{dog,3} = e_a \circledast \Phi \circledast e_{bit}$ } Trigrams

$bind_{dog,4} = \Phi \circledast e_{bit} \circledast e_{the}$

$bind_{dog,5} = e_a \circledast \Phi \circledast e_{bit} \circledast e_{the}$ } Quadgrams

$bind_{dog,6} = \Phi \circledast e_{bit} \circledast e_{the} \circledast e_{mailman}$

$bind_{dog,7} = e_a \circledast \Phi \circledast e_{bit} \circledast e_{the} \circledast e_{mailman}$ } Tetragram

where $\circledast$ denotes circular convolution, and Φ is a universal placeholder used in the computation of order information for every word in every position and sentence (i.e. constructed as a random vector in the same way as the environment vectors), such that $m_{dog} = m_{dog} + o_{dog}$.

In summary, BEAGLE uses the environment vectors to construct semantic memory vectors that represent the meaning of each word in the corpus as a combination of both context and order information. As the algebra indicates, the theory predicts that a word's meaning consists of its history of co-occurrence with, and position relative to, other words in sentences. Thus, BEAGLE implements and incorporates the wisdom from linguistics that, "You shall know a word by the company it keeps" (Firth, 1957; see also Hallet, 1967, for Wittgenstein's meaning as use).

In the context of this thesis and its aim to understand semantic distinctiveness, BEAGLE and LSA provide an articulate formal theory on how meaning emerges based on a history of

language experience. However, the theory is limited to describing word meaning, whereas the act of recognition requires additional assumptions about how words are stored in memory, retrieved from memory, and recognized.

Modeling Recognition Memory

There are several theories of recognition memory, all of which differ in important ways, but all of which are designed to model the same problem.

MINERVA 2 (Hintzman 1986, 1988) is a classic theory for human memory that has been used to model a number of phenomena, including decision making (Thomas, et al., 2008), categorization (Hintzman, 1986, 1988), cued recall (Hintzman, 1986), and most germane to the current thesis, recognition (Arndt & Hirshman, 1998; Clark, 1997). According to the theory, each new experience (e.g. each studied word) is encoded as a unique trace in memory. All retrieval is cue based. When a probe is presented to memory it activates all traces in parallel, where each trace becomes activated in proportion to its similarity to the probe. The sum of activation elicited by the probe quantifies the collective strength of response from memory, a value called the echo intensity. If the echo intensity is greater than a criterion, then the cue is identified as having been studied; else it is not. The decision mechanism is consistent with signal detection theory.

Formally, MINERVA 2 is also expressed as an algebraic theory. Memory is a matrix in which each row represents an event in the model's history. Encoding is assumed to be imperfect, an assumption that is represented by deleting some proportion of information in each trace (i.e. each value in an encoded trace is rewritten as a 0 with probability *L)*.

As for retrieval, all traces in memory are contacted in parallel at the same time. The extent to which traces are activated directly corresponds to how similar the probe is to the existing traces in memory:

$$S_i = \frac{\sum_{j=1}^{n} P_j M_{ij}}{\sqrt{\sum_{j=1}^{n} P_j^2} \sqrt{\sum_{j=1}^{n} C_j^2}}$$

where S_i is the similarity of the probe to trace i in memory, P_j is the j^{th} element of the probe, M_{ij} is the j^{th} element of the i^{th} row in memory, and n is the number of features that have non-zero elements in memory. Next, each trace is activated as a positively accelerated function of its similarity:

$$A_i = S_i^{\tau}$$

Where τ is an odd number, typically set to three, which increases the signal to noise ratio of the activated traces that are similar or dissimilar to the probe. The information retrieved from memory is a vector called the echo, C, that is a weighted sum of the traces in memory:

$$C_j = \sum_{i=1}^{m} A_i * M_{ij} \qquad \{\text{for each } j=1 \ldots n\}$$

where C_j is the j^{th} element of the echo, A_i is the magnitude of activation for the i^{th} trace in memory, M_{ij} is the j^{th} element of the i^{th} trace in memory, and m is the number of traces in memory. n is the number of elements, which in this case is the number of dimensions in each semantic vector.

Lastly, and specific to this project, how well a probe is recognized is determined by calculating the familiarity between a probe and the memory matrix:

$$f = \sum_{i=1}^{i=m} \left(\frac{\sum_{j=1}^{j=n} P_j \times M_{ij}}{\sqrt{\sum_{j=1}^{j=n} P_j^2} \sqrt{\sum_{j=1}^{j=n} M_{ij}^2}} \right)^{\tau}$$

where semantic vectors corresponding to words in the study list are stored in a memory matrix M, familiarity, f, for each word in the test set is computed relative to M, where n is the dimensionality of a representation, P_j is feature j in the probe, where τ is an odd number to

increase the signal to noise ratio, and M_{ij} is the j^{th} element in the i^{th} trace in memory. Familiarly, f, is then converted to a yes/no decision, d, relative to a criterion, $k = \bar{f}$:

$$d_i = \begin{Bmatrix} if\ f_i \geq \bar{f}, & then\ "yes" \\ if\ f_i < f, & then, "no" \end{Bmatrix}$$

Current Project

In the work that follows, I imported the semantic representations derived from BEAGLE and LSA into the MINERVA 2 theory of recognition memory. I then simulated recognition performance over the particular words in recognition memory experiments to predict performance for different categories of words (e.g. studied / not studied, distinctive / nondistinctive, and particular words). My intention was to show that the theory can predict recognition of distinctive words and can also predict the recognizability of words based on their distinctiveness as captured in the semantic vectors. To evaluate the theories, I conducted a series of five recognition memory experiments that manipulate the distinctiveness of words and the relationships between unstudied foils presented at test.

Semantic Representations

As explained at the outset of this thesis, my aim was to conduct experiments in service of testing a computational account of memory. To conduct that analysis for Experiment 1, and in the rest of the experiments that follow, I defined word meaning using two distributional theories of semantics (BEAGLE and LSA) and then simulated recognition memory performance with an instance-based theory of semantics (MINERVA 2) based on those representations. My approach differs from the classical assumptions of recognition models in which words are represented by random vectors consisting of binary values of either +1's or -1's such that all words are assumed to be unrelated in meaning, but is consistent with more recent work that assess vectors from distributional models of semantics in an effort to predict recognition and conditional on word

relationships based on a history of reading and language experience (Chubala et al., 2016; Johns, Jones, & Mewhort, 2019).

Although there are a variety of models that could be used to derive semantic representations (e.g. Word2Vec or GloVe), BEAGLE and LSA provide convenient and cognitively valid methods to derive semantic vectors. By comparison, Word2Vec (Mikolov et al., 2013) and GloVe (Penningston, Socher, & Manning, 2014) follow more from issues and ideas from the fields of machine learning and computational linguistics.

Thus, BEAGLE and LSA were used to derive semantic vectors from the TASA corpus (Touchstone Applied Science Associates, Inc.). TASA was used as it is a common corpus used for the development of semantic vectors and contains texts that cover a broad range of educational topics. The vectors that were derived were 2000 dimensions for BEAGLE and 300 dimensions for LSA. (Jones & Mewhort, 2007; Recchia, et al., 2015; Landauer & Dumais, 1997).[1] Although BEAGLE vectors can include both context and order information, I used only the context vectors for construction of the semantic vectors for BEAGLE. Theoretical motivation for this was that syntax information derived from the order vectors is unimportant for deriving meaning (see Jones & Mewhort, 2007). In addition, as a practical advantage, using only the context vectors is computationally more efficient.

Of course, it is difficult to visualize relationships in 4+ geometric space. However, once these semantic vectors are derived, a dimension reduction technique called multidimensional scaling (MDS) can be applied to visualize and plot words in in a 2- or 3-dimensional space instead of the 2000 or 300 dimensional spaces in which words are embedded in BEAGLE and LSA respectively (see Shepard 1964, 1987). In practical terms, words that are similar to one

[1] The matrix of word vectors derived by LSA that was used in all experiments is available as an .rda file at http://www.lingexp.uni-tuebingen.de/z2/LSA spaces/

another are plotted close together in a semantic space, whereas words that are dissimilar in meaning are plotted further apart. For example, in a list of countries with a semantic outlier such as *cobra* embedded, one can see the countries clustered close together, with the word *cobra* plotted further away (see Figure 2 for an example). For comprehension purposes, this is how semantic distinctiveness will be defined, although computationally, the relationships between words will be defined as the cosine angle between the vectors representing those words in the 2000-dimensional space for BEAGLE and the complete 300-dimensional space for LSA.

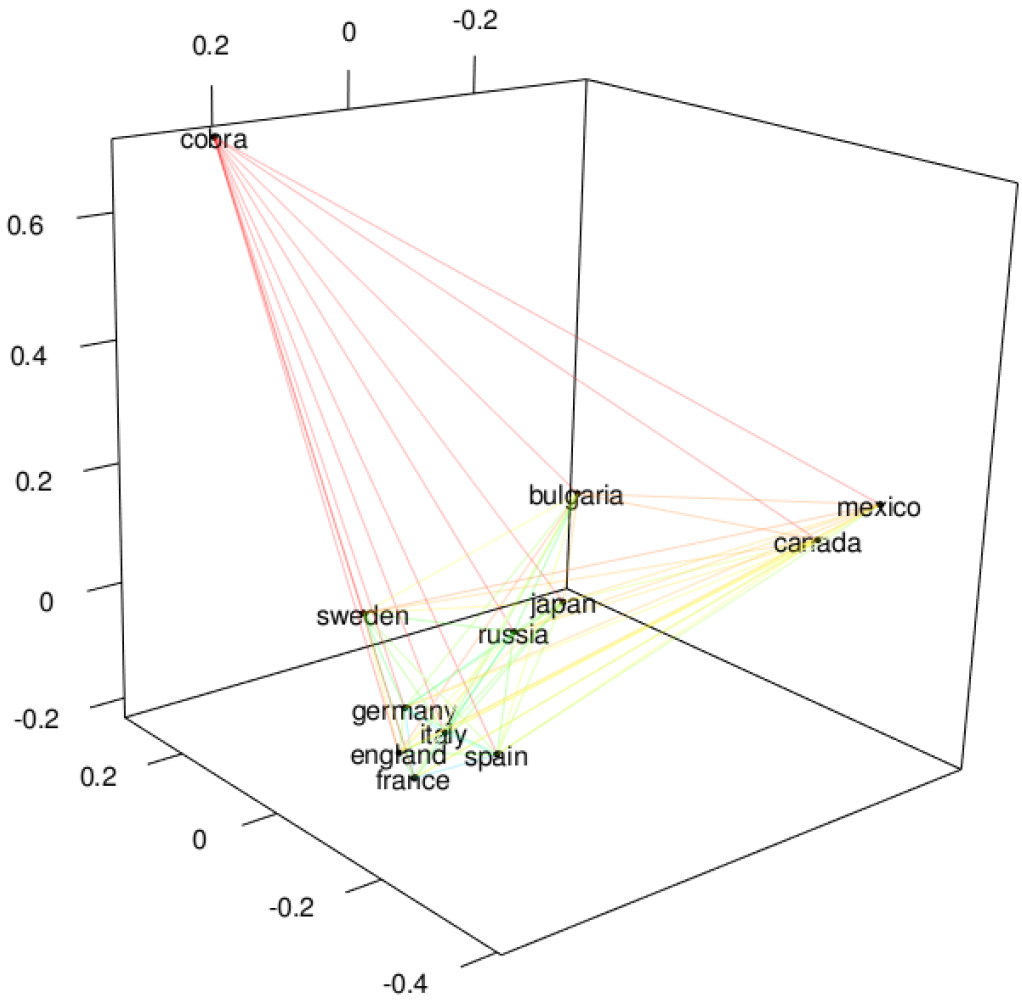

Figure 2. A multidimensional scaling solution and graphical representation of a 3-dimensional semantic space (as opposed to 2000) derived using BEAGLE. Semantically similar words (countries) are close together, whereas the semantic dissimilar word (cobra) is further away. The luminance of line colors represents the magnitude of cosine values between two words, with darker colors representing lower cosine values and lighter colors representing higher cosine values.

To simulate, study semantic vectors that correspond to words in the study list were stored in a memory matrix, *M,* with each dimension stored with probability *L* (else encoded zero). As imperfect encoding is assumed with the learning rate parameter *L*, each simulation will yield different results. Therefore, I conducted 1000 independent simulations with *L* set to 0.05 for LSA and *L* set to 0.025 for BEAGLE in Experiment 1. For both simulations, τ in MINERVA 2 was set to 7 (i.e. a sharp signal to noise ratio over trace activation).

Familiarity, *f,* for each word in the test set was computed relative to *M* and then converted to a yes/no decision, *d*, relative to a median criterion $k = f$.

$$d_i = \text{if } f_i \geq \bar{f}, \text{ then "}yes\text{", if } f_i < f, \text{ then "}no\text{".}$$

This familiarity will be stored over each subsequent simulation and the resulting aggregate for each word will serve as an average yes/no response to each probe in the test list, allowing results from the model to be compared to the obtained experimental results. Performance of the model and the experimental results will be compared using linear regression and computing the proportion of variance in participants' yes/no decision rates over all items by R^2.[2]

Experiment 1

The theories presented in this thesis provide predictions for recognition in a way that is conditional on semantic relationships. As such and as previously stated, the models ought to provide predictions that define semantic distinctiveness in this way. Semantically similar words

[2] R^2 was used instead of RMSE, BIC or other methods as the goal was to measure the relative fit of the models, as comparing the number of parameters between LSA and BEAGLE is theoretically problematic.

are represented closer together in a semantic space and semantically dissimilar words are represented as farther apart. With these models serving as a theory of distinctiveness in this way, they ought to provide predictions that are consistent with other accounts of distinctiveness. That is, semantically dissimilar words ought to "stick out" among a list of semantically related words. As an initial test of the theories, Experiment 1 was conducted.

Experiment 1 was a standard yes/no recognition memory test, where participants studied 8 blocks of words. Each block of words was defined by a particular category, where one word was a deliberate semantic outlier, deeming it distinctive. For example, if presented with a list of occupations, the semantic outlier was a word like cobra. Following the study phase, participants were tested for recognition of the words they studied (i.e., the category targets and distinct targets) as well as unstudied words. The unstudied words were of three kinds: words that people did not study and that were unrelated to the studied list, words that people did not study but that were related to the studied categories, and words that people did not study but that were related to the studied distinctive words (see Figure 1 for an example of a list). Further, if the theory could predict recognition for classes of words, can it also predict recognition of particular words? If successful the theory would not only account for breadth, but precision as well.

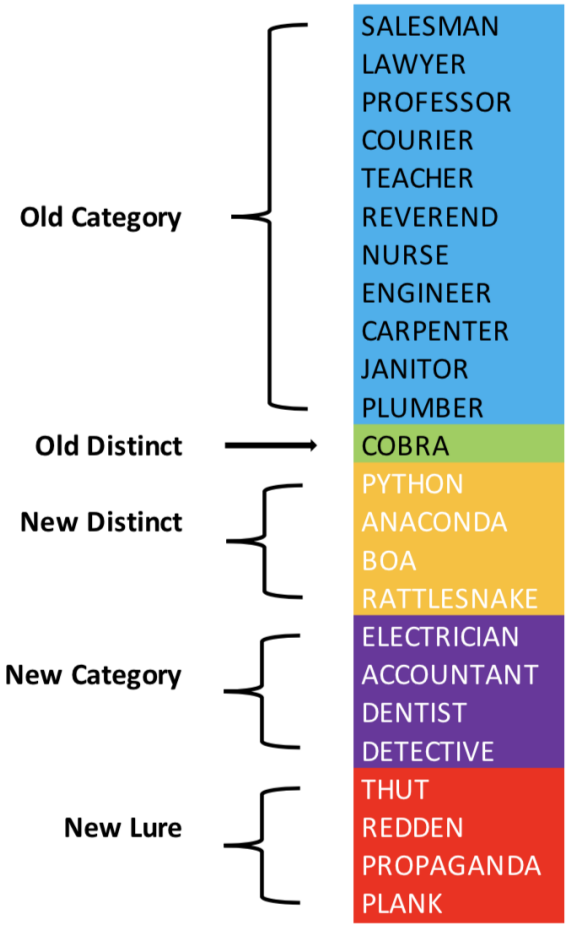

Figure 1. A graphic displaying the construction of a list, categorizing the different types of study words and new test lures.

This experiment is similar to a study conducted by DeSoto and Roediger (2014). In that experiment, they constructed categorized study lists (e.g. birds) for participants to study and, then, tested them with a yes/no recognition test for the previously presented categorized study lists, new words that were related to the studied categories, and new words that were unrelated taken from new categories. They found that participants responded to old studied words more often than new related words, and more often new related words than new unrelated words. The experiment conducted here is extremely similar, however with one crucial difference: during the study phase, categorized lists had one critical-distinctive word embedded within them. Although

the minor change, this design offers a convenient way to examine semantic relatedness and thus semantic distinctiveness. Moreover, it is grounded in a rich theoretical history, and a history of empirical success (Dewhurst, 2001; Dewhurst & Anderson, 1999).

On the broadest level, the first research question that was posed in Experiment 1 was whether people could recognize studied words better than words they did not study. The prediction that followed was that indeed, participants would be able to distinguish studied words from unstudied words, consistent with most recognition experiments. Secondly, predicted was that participants would recognize the distinctive words better than the category words that they studied. This would be consistent with a classic von Restorff effect.

As for the false alarm rates to the new related and unrelated words presented at test, predictions were that participants will have trouble rejecting words related to the category words the most, consistent with a classic Deese-Roediger-McDerott (DRM) effect (Deese, 1959; Roediger & McDermott, 1995), followed by new words that are related to the studied distinct words, consistent with a distinctiveness effect, and false alarm rates being the lowest for new unrelated words, again consistent with most recognition experiments.[3] However, before the empirical results are presented, the simulation results from the two theories will be presented.

Method

Participants

Seventy-five participants were recruited from the University of Manitoba SONA psychology participant pool. Participants were run in groups of up to eight at a time, with the

[3] A DRM effect refers to the result obtained from the popular DRM paradigm, which consists of presenting participants with a list of related words (e.g. *bed, rest, dream*) that are related to a critical word that is not presented (e.g. *sleep*), resulting is high rates of false recall (or recognition) of the unstudied critical word.

procedure typically lasting no longer than 15 minutes. Participants received 1 credit towards completion of their introduction to psychology course.

Apparatus

Participants were seated in front of Dell OptiPlex PCs, all equipped with QWERTY keyboards and 22" LCD monitors. All response data were collected by keyboard input. Presentation of and responses to stimuli within the experiment was conducted using PsychoPy3 software (Garaizar & Vadillo, 2014; Peirce, 2007, 2008; Peirce & MacAskill, 2018).

Materials

The selection of words was constrained to words in the TASA corpus, to ensure that all words in the experiment had corresponding vectors for simulating performance in the computational modelling analysis. Again, the TASA corpus contains text that covers a broad range of educational topics. The materials included 96 study words and 96 test words. The 96 study words were broken down into eight categorized lists, consisting of 11 words per category, with 1 unrelated distinctive word embedded in each list. These were taken from common taxonomic categories consisting of high frequency nouns, such as a list of different occupations or sports. A complete list of the stimuli for Experiment 1 appears in the Appendix. Multiword category exemplars were avoided (e.g., New York).

Along with the previously studied test words that were presented to participants, the 96 new test words were composed of 3 kinds of words: 32 were new words that were related to the distinct study words, 32 were new words that were related to the studied category words, and 32 were new words that were unrelated to the study words and other unstudied words.

Procedure

At the beginning of the experiment, subjects were presented with a welcome screen that read: “Please read the consent form and instructions. When you are finished the experimenter will answer any questions you may have and then get you started”. The participants were not able to begin the experiment until the experimenter pressed the letter ‘e’ on the keyboard.

At study, the 96 study words were presented from the 8 categorized lists of 11 words per category and 1 unrelated distinctive word. At test, subjects made yes/no judgements for 192 test words including the 88 old category words, the 8 old distinctive words, the 32 new distinct related words, the 32 new category related words, and the 32 new unrelated words.

These categorized lists of words were presented in a randomized blocked format, with 12 trials within each block that were also randomized, with the entire study phase consisting of 8 blocks in total. Each word was presented one at a time, and appeared on the screen for 2000 ms, with a 500 ms break in between each trial during which time the screen was blank. All stimuli were presented on the computer screen in white uppercase Arial font, with letters that were 1cm tall, with a dark grey background, in the center of the screen. All stimuli were easily visible and readable. Participants were then presented with a rest screen in between each study block that read, “Take a break if you need. When you are ready, press the spacebar to continue”.

Testing occurred after the study of all blocks. Although the study words and categories were presented in a blocked format, the test words were not. The 192 test words were presented in a completely randomized fashion. The rationale for this choice was that MINERVA 2 cannot account for blocked testing. Following the study phase, the participants were presented with a screen that read, “Take a rest if you need. Next, your job will be to indicate whether you previously studied a word by pressing 'y', otherwise press 'n'. Keep your index fingers on these

keys. It is important that you respond accurately but also quickly.[4] When you are ready press "i" to continue". When the participant pressed "i" the screen was cleared and 500 ms later the first test word was presented. After the participant responded by pressing the "y" or "n" key, the screen was cleared for 500 ms after which the next word was presented. When the participant finished, they were presented with a screen thanking them for their participation and that instructed them to see the experimenter for a debriefing form.

Computational Simulations

Simulation results using the LSA and BEAGLE semantic vectors coupled with MINERVA 2 are shown in Figure 3. The first column shows item recognition rates for the empirical data, item recognition estimates derived from LSA, and item recognition estimates derived from BEAGLE from top to bottom, respectively. The second column shows recognition rates for words from their respective word classes, for the empirical data and the simulated LSA and BEAGLE results from top to bottom, respectively. The simulation results from LSA and BEAGLE provided the same pattern of predictions that were derived intuitively based on previous research, except for the von Restorff effect. LSA nor BEAGLE predicted that the old distinct words should be better recognized than the old category words. Therefore, this pattern of results will also be predicted for Experiment 1 to be consistent with the theories.

[4] Although the participants were instructed to respond quickly, response time data was not analyzed for the purposes of this project.

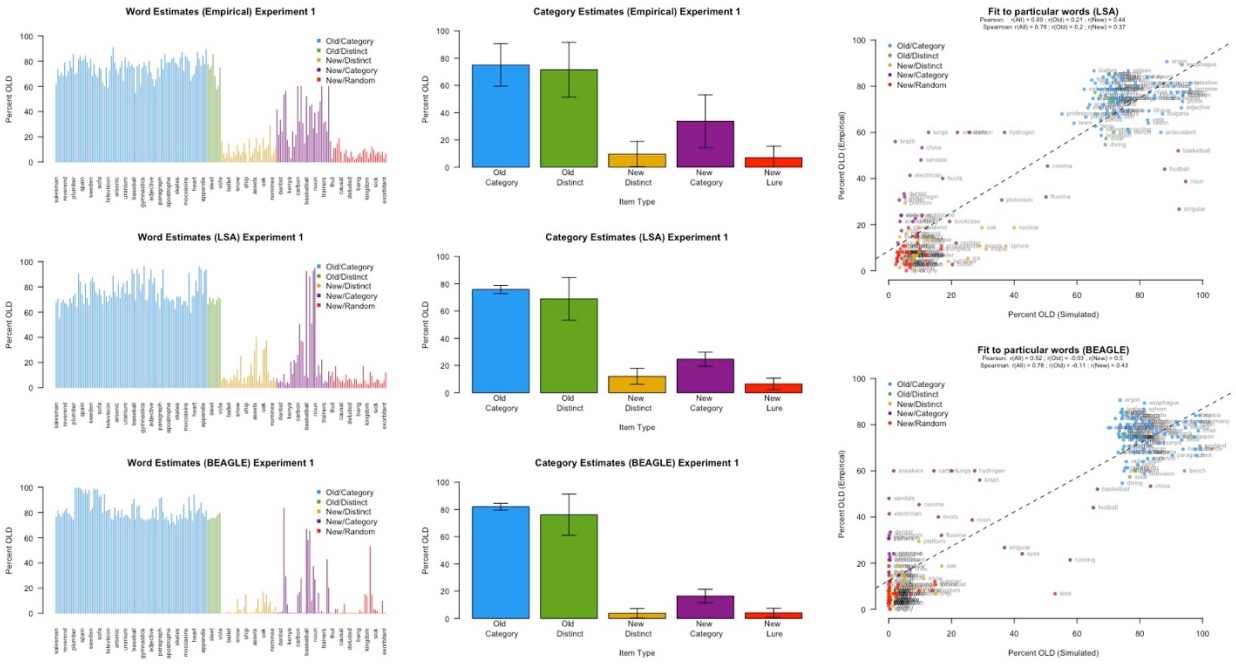

Figure 3. Columns 1 and 2 show item and categorical results for the empirical, LSA and BEAGLE results respectively in Experiment 1. Error bars represent ± 1 SD. Column 3 shows scatter plots with best fitting regression lines for LSA and BEAGLE in comparison to empirical data. Each point represents an individual word.

More specifically, between the different classes of words, both models predicted that subjects ought to respond "yes" more frequently to old words more than new words, where old category words had a higher number of "yes" responses than the old distinct words. Within the new classes of words, both LSA and BEAGLE provided predictions that proposed that the highest proportion of "yes" responses (or false alarms) ought to be for the new category words, followed by the new distinct words, and lastly followed by the new unrelated words. Although LSA and BEAGLE provided similar categorical predictions, when examining the item estimates, the predictions from both models were quite different.

Both models predicted that some words ought to be better recognized than others. When examining the item specific predictions, we can first see that there is far more variance among the new classes of words than there is among the old classes of words. Within those classes, both

theories exhibited more stable predictions among the old classes of words, however both LSA and BEAGLE exhibited much more variance both within and between the new classes of words. For example, within the new category class of words, LSA predicted that the word *singular* ought to have a very high number of "yes" responses elicited, but a very low number of "yes" responses for the word *brazil.* As for BEAGLE, within the new category class of words, the theory predicted that participants ought to false alarm to the word *running*, but that they ought not to the word *sneakers*. Therefore, even in the case of the strongest predictions that LSA and BEAGLE derived, there were differences between the theories. Thus, this provided a critical test for both theories. Lastly, it should be noted that LSA provided more variance among item estimates than BEAGLE did. Of course, with both theories providing such strong predictions, this created a falsifiable test of item specific recognition rates.

Results

The top row of Table 1 shows the participants' mean percentage of "yes" responses to the five different classes of test words. A repeated measures ANOVA revealed statistically significant differences in "yes" responses as a function of item type, $F(4, 74) = 441.00$, $p < 0.01$, $MSE = 545.00$, $\eta^2 = 0.86$. In terms of the planned comparisons and specific differences between the different classes of words at the most general level, participants responded "yes" to studied words ($M = 74.75$) more frequently than they did to words they did not study ($M = 16.72$), $t(74) = 28.88$, $p < 0.01$. Within the studied words, subjects did not respond "yes" more frequently to the old distinct words ($M = 71.50$) than to the old categorized words ($M = 75.05$), $t(74) = 1.62$, $p = 0.11$. Between the studied and unstudied words, subjects responded "yes" more frequently to old category words ($M = 75.05$) than to new category words ($M = 33.67$), $t(74) = 20.05$, $p < 0.01$ and responded "yes" more frequently to old distinct words ($M = 71.50$) than to new distinct

words (M = 9.58), $t(74) = 23.92$, $p < 0.01$. Within the unstudied words, participants responded "yes" more frequently to new category words (M = 33.67) than to new distinct words (M = 9.58), $t(74) = 13.79$, $p < 0.01$, and in terms of old category and new category words combined, versus old distinct and new distinct words combined, participants responded "yes" more frequently to the categorical classes of words (M = 64.01) than to the distinct classes of words (M = 21.97), $t(74) = 9.30$, $p < 0.01$. Lastly, the two new classes of words that were related to previously studied words (the new category and new distinct words) had significantly higher false alarm rates (M = 21.63) than the new unrelated lures that were unrelated to all of the other words (M = 6.92), $t(74) = 11.05$, $p < 0.01$.

Table 1. *Percentage "Yes" Responses as a Function of Test Item Type*

	Old		New		
	Category	Distinct	Category	Distinct	Unrelated
Experiment					
Experiment 1	75.05 (15.58)	71.50 (20.09)	33.67 (19.44)	9.58 (9.25)	6.92 (8.48)
Experiment 2	68.75 (12.78)	56.82 (21.03)	20.60 (11.07)	8.81 (6.59)	8.95 (9.46)
Experiment 3	60.43 (16.60)	57.20 (24.29)	25.19 (15.45)	17.46 (15.90)	16.79 (17.03)
Experiment 4	60.55 (17.32)	64.32 (21.68)	25.70 (18.20)	19.66 (16.95)	16.66 (15.15)

Note. Values in parentheses are standard deviations.

Computational Comparisons

Once again, the first column shows item recognition rates for the empirical data, item recognition estimates derived from LSA, and item recognition estimates derived from BEAGLE from top to bottom, respectively. The second column shows recognition rates for words from their respective word classes, for the empirical data and the simulated LSA and BEAGLE results from top to bottom, respectively. Lastly, the third column shows the relationship between the empirical and simulated item recognition rates for both LSA and BEAGLE (top to bottom,

respectively). The scatter plots show a best fitting regression line, with values for Pearson (r_p^2) and Spearman rank-order (r_s^2) correlations. Further correlations are calculated for hits and false alarms separately, as they appear as two distinct clusters.

On the categorical and item levels, both LSA and BEAGLE exhibited quite similar patterns of results to one another and to the empirical data. Specifically, when looking at the particular word recognition rates in column 1, it is evident that both models do a pretty good job at predicting item specific hit and false alarm rates compared to the empirical data. However, there are some differences. For example, both LSA and BEAGLE under predict the rate of false alarm rates for new category words, and BEAGLE under predicts the false alarm rates of the new distinct words and new lures.

When comparing the results of LSA and BEAGLE on the item level with the empirical data, we can see the relationships plotted in the third column of Figure 3. The scatter plots here show how well the models predict that people ought to recognize particular words (on the x-axis) relative to how well they did recognize those particular words (on the y-axis). Firstly, ignoring the distinction between old and new test words, there was a strong linear trend for LSA with correlations of $r_p(190) = 0.89$, $p < 0.01$, $r_s(190) = 0.79$, $p < 0.01$, $r_p^2 = 0.79$, and the shared variance of the ranked data $r_s^2 = 0.62$. However, as there are two distinct clusters within the scatter plot, one representing hits in the upper right quadrant, and the other representing false alarms in the lower left quadrant, these measurements exaggerate and possibly distort the quantitative model fit. Thus, correlations were calculated separately for these two distinct clusters. The resulting correlations were a much reduced $r_p(94) = 0.21$, $p = 0.04$, $r_s(94) = 0.21$, $p = 0.04$, $r_p^2 = 0.04$, and the shared variance of the ranked data $r_s^2 = 0.04$, for hits, where the

resulting correlations for false alarms were $r_p(94) = 0.45$, $p < 0.01$, $r_s(94) = 0.38$, $p < 0.01$, $r_p^2 = 0.20$, and the shared variance of the ranked data $r_s^2 = 0.14$ for false alarms.

The results using BEAGLE were $r_p(190) = 0.88$, $p < 0.01$, $r_s(190) = 0.80$, $p < 0.01$, $r_p^2 = 0.77$, $p < 0.01$ and the shared variance of the ranked data $r_s^2 = 0.64$, $p < 0.01$ when ignoring the distinction between old and new test items, were $r_p(94) = 0.25$, $p = 0.01$, $r_s(94) = 0.18$, $p = 0.07$, $r_p^2 < 0.06$, and the shared variance of the ranked data $r_s^2 = 0.03$ for hits alone, and were $r_p(94) = 0.62$ $p < 0.01$, $r_s(94) = 0.38$, $p < 0.01$, $r_p^2 = 0.38$, and the shared variance of the ranked data $r_s^2 = 0.14$ for false alarms alone.

The results from Experiment 1 were consistent with predictions derived from LSA and BEAGLE on the categorical level. Though I a priori predicted, based on intuitions, that old distinctive words would be better recognized than old category words, I did not observe this result, nor did the theory predict this result. Intriguingly, both models predicted the same result in Experiment 1, contrary to my intuitions. Further, when evaluating the relationship between simulated results and empirical results ignoring the distinction between old and new items, both models had strong linear trends. However, when tracking hits and false alarms separately, the corresponding relationships weaken, with both models tracking the item specific false alarms better than item specific hits.

Taken together, results from Experiment 1 provide evidence that LSA and BEAGLE do a serviceable but not a particularly good job at tracking participants' recognition performance. Although LSA outperformed BEAGLE for overall performance when ignoring the distinction between old and new items, neither LSA or BEAGLE differed significantly in this performance nor did they differ when tracking hits and false alarms separately at the $p < 0.05$ significance level. In sum, both models provided good predictions at the category level, however that was not

the case at the item specific level. Even for predictions of words in the strongest cases, (e.g. high and low false alarm rates for words like *singular* and *brazil* in LSA, and high and low false alarm rates for words like *running* and *sneakers* in BEAGLE), the empirical results were not consistent with these predictions.

Experiment 2

Given that MINERVA 2 does not predict that there should be a difference in regards to whether the words are presented in a blocked or unblocked order, Experiment 2 was conducted to test if the models predict this result solely empirically. Therefore, words were presented in a random rather than blocked order. If the models are correct that study order does not matter, then the results of Experiment 2 should be consistent with the result of Experiment 1. As such, predictions were that the results in Experiment 2 ought to be consistent with those found in Experiment 1.

Participants

Twenty-two participants were recruited from the University of Manitoba subject pool as before. However, no participant who was a part of Experiment 1 was eligible to participate in Experiment 2, and this holds true for the rest of the experiments that follow.

Apparatus

Apparatus was identical to that in Experiment 1.

Materials

Materials were identical to those in Experiment 1.

Procedure

Experiment 2 followed the same procedure as Experiment 1 except that the words were presented in a random rather than blocked order and there were no breaks presented during the study phase.

Results

The second row of Table 1 shows the mean percentages of "yes" responses to the five different classes of test words in Experiment 2. A repeated measures ANOVA revealed statistically significant differences in "yes" responses as a function of item type, $F(4, 21) = 187.41$, $p < 0.01$, $MSE = 489.90$, $\eta^2 = 0.90$. In terms of the planned comparisons and specific differences between the different classes of words, participants responded "yes" to studied words ($M = 67.76$) more frequently than they did to unstudied words ($M = 12.79$), $t(21) = 21.55$, $p < 0.01$. Within the studied words, subjects did not respond "yes" more frequently to the old distinct words ($M = 56.82$) than to the old categorized words ($M = 68.75$), $t(21) = 3.17$, $p < 0.05$. Between the studied and unstudied words, subjects responded "yes" more frequently to old category words ($M = 68.75$) than to new category words ($M = 20.60$), $t(21) = 25.08$, $p < 0.01$ and responded "yes" more frequently to old distinct words ($M = 56.82$) than to new distinct words ($M = 8.81$), $t(21) = 11.95$, $p < 0.01$. Within the unstudied words, participants responded "yes", or false alarmed more frequently to new category words ($M = 20.60$) than to new distinct words ($M = 8.81$), $t(21) = 6.11$, $p < 0.05$, and in terms of old category and new category words combined, versus old distinct and new distinct words combined, participants did not respond "yes" more frequently to the two categorical classes of words ($M = 55.91$) than the distinct classes of words ($M = 18.41$), $t(21) = 0.03$, $p = 0.97$. Lastly, in comparison to Experiment 1, the new unrelated lures were better recognized ($M = 8.95$) than the two new classes of words ($M = 14.71$) that were

related to previously studied words (the new category and new distinct words), $t(21) = 3.89$, $p < 0.01$.

Aside from three sets of comparisons, the results in Experiment 2 were consistent with predictions and Experiment 1. The first comparison that was inconsistent with Experiment 1 was the contrast of old and new categorical classes of words versus the old and new distinct classes of words. Contrary to Experiment 1, this difference was not statistically significant in Experiment 2. However more interestingly, the comparison between old category words and old distinct words was statistically significant, where old category words had a significantly higher number of "yes" responses than the old distinct words. This could be thought to constitute a non-von Restorff effect, and is addressed in the general discussion. Lastly, unlike Experiment 1, the new unrelated lures had a higher number of "yes" responses than the other two new classes of words that were related to the previously studied words.

Discussion of Experiments 1 and 2

As an initial empirical demonstration, Experiment 1 was conducted in a blocked format during the study phase such that it emphasized the distinctiveness of the critical words. However, due to the models' assumption of path independence, it should not matter what order or structure the words are presented in during the study phase. As such, Experiment 2 was conducted to test that assumption empirically.

In both experiments old words were better recognized than new words, with no significant differences between old category and old distinct words in Experiment 1. In Experiment 2, old category words were reliably recognized better than old distinct words. When it came to new words or foils, people false alarmed to new distinct and new category words (that is, new words related to study words) than to new unrelated lures, where specifically, there were

significantly more false alarms to the category related lures than to the distinct related lures. Both the LSA and BEAGLE semantic vectors coupled with MINERVA 2 predicted the same general pattern of results. Further, when it came to item specific recognition in Experiment 1, LSA and BEAGLE also did a poor job at matching people's recognition of the old/studied words but a competent job at matching people's performance of recognizing new/unstudied words.

Experiments 1 and 2 provide evidence that the relationship between categorized words at study coupled with semantic outliers, and related lures to both of these classes of words presented at test (along with random unrelated lures) affects recognition performance, whether words are presented in random or blocked fashion at study.

Another popular procedure closely related to the above one used in Experiments 1 and 2 is the DRM protocol. Pioneered by Deese in 1959, and popularized by Roediger, and McDermott in 1995, the procedure has had years of successful and repeated demonstrations in a variety of different areas of memory research. In a classic DRM procedure, a number of related category exemplars are presented during a study phase (e.g. *bed*, *rest*, *dream*) where one strongly related theme word is absent during study but presented at test (e.g. *sleep*). False alarm rates to this new critical word are elevated and sometimes equal in magnitude to hit rates for old words that were in fact presented during the study phase. The motivation and purpose of the next 2 experiments was to extend the demonstrations from Experiments 1 and 2 into a framing of DRM research.

Experiment 3

Using a selection of the materials from the Appendix of Stadler, Roediger and McDermott (1999), the procedure used in Experiment 2 was extended using DRM lists. As there is a rich history within the realm of DRM research, Experiment 3 was conducted to extend this approach using these materials using the same experimental procedure as Experiment 2.

Method

Participants

Ninety-nine participants were recruited from the University of Manitoba SONA subject pool. This sample size was deliberately large to support both categorical comparisons among test item type as before, but also measure and compare item recognition rates like in Experiment 1.

Apparatus

The apparatus in Experiment 3 was identical to those in Experiments 1 and 2.

Materials

Materials were taken from the Appendix of Stadler, Roediger, & McDermott (1999). Of the 35 lists provided in their Appendix, eight were randomly selected for use in this experiment. These lists were comprised of 15 words, where each word within a list was related to a critical target. For example, for a list that had the critical target as *anger*, DRM words that would be included in this list would be: *mad, fear, hate, rage*, etc. The DRM lists that were randomly selected and used in this experiment were: MUSIC, THIEF, WINDOW, KING, SWEET, HIGH, FOOT, and CITY, where the main category exemplar (e.g. *anger*) was omitted as a category member.

To be consistent with the previous experiments, 11 of the 15 words in each list were used as the studied category words and 4 were used as the new unstudied words (lures) that were related to those 11 studied category words. Studied critical words, and unstudied lures related to these critical words, along with the unstudied lures that were unrelated to all studied and unstudied words were systematically selected using LSA. Lure selection was restricted to a list of high frequency nouns that were taken from the Toronto Word Pool and a selection of nouns taken from the MRC database with Kucera-Francis frequency values between 40 and 70 000

(Friendly, 1982; Coltheart, 1981; Wilson, 1988). After removing any duplicated words or multi-word exemplars (e.g. New York), the resulting pool consisted of 1996 words. Selection was done in an iterative process, where critical distinct words were selected first, where words related to these distinct critical words were selected next, and unrelated new words having no relationship to any of the studied or unstudied words were selected last. As before, if there were multiword exemplars that were included in these lists, they were excluded and the next category word was chosen. A complete list of the materials that were used in Experiment 3 appear in the Appendix.

Procedure

The procedure in Experiment 3 was identical to that in Experiment 2.

Computational Simulations

Simulation results using the LSA and BEAGLE semantic vectors coupled with MINERVA 2 for Experiment 3 are shown in Figure 4.[5] One thousand independent simulations were conducted with *L* set to 0.05 for LSA and *L* set to 0.016 for BEAGLE. For both simulations, τ in MINERVA 2 was set to 7. The first column shows item recognition rates for the empirical data and item recognition estimates derived from LSA and BEAGLE. The second column takes these item estimates and collapses them into their respective word classes, again for the empirical data and the simulated LSA and BEAGLE results.

[5] The word 'heart' was inadvertently selected twice as a lure, once as a category related lure and once as a distinct related lure. As such it was omitted from computational simulations for both BEAGLE and LSA as well as from the empirical results.

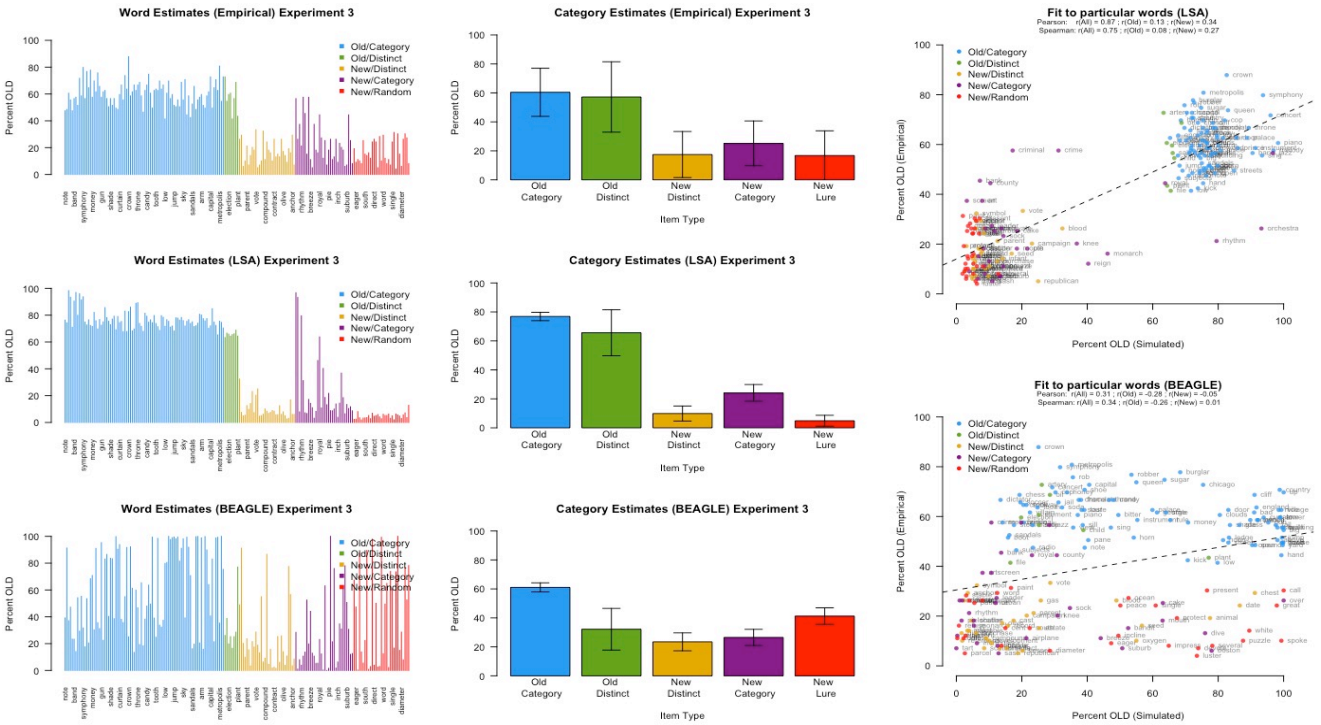

Figure 4. Columns 1 and 2 show item and categorical results for the empirical, LSA and BEAGLE results respectively, in Experiment 3. Error bars represent ± 1 SD. Column 3 shows scatter plots with best fitting regression lines for LSA and BEAGLE in comparison to empirical data. Each point represents an individual word.

Firstly, the simulation results from LSA and BEAGLE did not provide the same pattern of predictions in Experiment 3. On the categorical level, the simulation results from LSA offered predictions that were the same general pattern that were exhibited in Experiment 1. However, simulation results from BEAGLE offered a pattern of predictions that were vastly different, particularly when it came to old distinct words, and false alarm rates for the new related and unrelated lures. As compared to the predictions derived by LSA, BEAGLE predicted that recognition rates for old distinct words ought to be quite a bit lower, and that there ought to be much higher false alarm rates for the new distinct words and new unrelated words.

As for the item specific predictions, LSA again predicted differences in recognition rates that were more subtle between specific words for the old classes of words, but far more dramatic

for the false alarm rates among new words. Specifically, within the new category class of words, some of the predictions for false alarm rates were at the same magnitude for that of hit rates for old words, and some were at the same magnitude for that of false alarm rates for the new unrelated words.

However, the predictions derived by BEAGLE are entirely inconsistent with the variance exhibited across different classes of words that was predicted by LSA. There are dramatic differences in predictions of recognition rates across all five classes of words, which unlike LSA, are not contingent on whether words are old or new. For BEAGLE, there is a large amount of variance predicted both between and within the different classes of words. Both theories again predicted that some words ought to be better recognized than others. For example, within the new category class of words, LSA predicted that the word *orchestra* ought to have a very high number of "yes" responses elicited, but a very low number of "yes" responses for the word *criminal.* As for BEAGLE, within the new category class of words, the theory predicted that participants ought to false alarm to the word *boston*, but that they ought not to the word *crime*. Therefore, even in the case of the strongest predictions that LSA and BEAGLE derived, there were differences between the theories. Thus, this again provided a critical test for both theories.

Given these differences, Experiment 3 provided an opportunity to test which model did a better job at tracking participants' recognition behavior, both on the categorical and item level. Further, as LSA was used to derive the related and unrelated materials to the DRM materials used in Experiment 3, this also created an opportunity to examine how well LSA did in producing variance among selected items that may be observed among human participants. Finally, a critic may argue that LSA and BEAGLE are not equivalent theories, thus it may have

been possible to find a list of words (such as in Experiment 3) that give two different sets of predictions.

Results

The third row of Table 1 shows the mean percentage of "yes" responses to the five different classes of words. A repeated measures ANOVA revealed statistically significant differences in "yes" responses as a function of item type, $F(4, 98) = 226.20$, $p < 0.01$, $MSE = 949.90$, $\eta^2 = 0.73$. A planned comparison between the different classes of words, confirmed that participants responded "yes" to studied words ($M = 60.16$) more frequently than they did to unstudied words ($M = 19.81$), $t(98) = 20.34$, $p < 0.01$. Within the studied words, subjects did not respond "yes" significantly more frequently to the old distinct words ($M = 57.20$) than to the old categorized words ($M = 60.43$), $t(98) = 1.70$, $p = 0.09$. Between the studied and unstudied words, subjects responded "yes" more frequently to old category words ($M = 60.43$) than to new category words ($M = 25.19$), $t(98) = 21.48$, $p < 0.01$ and responded "yes" more frequently to old distinct words ($M = 57.20$) than to new distinct words ($M = 17.46$), $t(98) = 15.16$, $p < 0.01$. Within the unstudied words, participants responded "yes", or false alarmed more frequently to new category words ($M = 25.19$) than to new distinct words ($M = 17.46$), $t(98) = 8.53$, $p < 0.01$, and in terms of old category and new category words combined, versus old distinct and new distinct words combined, participants responded "yes" more frequently to the two categorical classes of words ($M = 51.03$) than the two distinct classes of words ($M = 25.41$), $t(98) = 2.17$, $p = 0.03$. Lastly, the two new classes of words that were related to the studied items ($M = 21.33$) were better recognized than the new unrelated lures ($M = 16.79$) that were related to previously studied words (the new category and new distinct words) $t(98) = 5.82$, $p < 0.01$.

Computational Comparisons

Simulation results using the LSA and BEAGLE semantic vectors coupled with MINERVA 2 for Experiment 3 are shown in Figure 4. One thousand independent simulations were conducted with *L* set to 0.05 for LSA and *L* set to 0.016 for BEAGLE. For both simulations, τ in MINERVA 2 was set to 7. The first column shows item recognition rates for the empirical data and item recognition estimates derived from LSA and BEAGLE. The second column takes these item estimates and collapses them into their respective word classes, again for the empirical data and the simulated LSA and BEAGLE results. The third column shows the relationship between the empirical results and the predicted results for both LSA and BEAGLE.

The scatter plots show a best fitting regression line, with values for Pearson (r_p^2) and Spearman rank-order (r_s^2) correlations. Further correlations are calculated for hits and false alarms separately, as they appear as two distinct clusters.

On the categorical and item levels, LSA did not have the same pattern of results as the empirical data, as shown in columns 1 and 2 of Figure 4. Specifically, when looking at the particular word recognition rates in column 1, it is evident that LSA did not do a good job at predicting the item specific hit and false alarm rates compared to the empirical data. When looking at the simulation results for BEAGLE, the pattern of results are wholly inconsistent with the empirical facts. Hit rates for old category words were not comparable to those in the empirical data, and old distinct words were underpredicted. False alarm rates for new distinct words and new category words are also different from the empirical data, and the rate of false alarms for the new lures is far beyond those present in the empirical data. This is shown in both the word estimates in the lower left-hand corner of Figure 4, as well as in the category estimates in the middle of the last row in Figure 4.

The scatter plots here again show how well the models predict that people

ought to recognize particular words on the (x-axis) relative to how well they did recognize those words (empirical data on the y-axis). Ignoring the distinction between old and new words, there was a linear trend for LSA with $r_p(188) = 0.87$, $p < 0.05$, $r_s(188) = 0.75$, $p < 0.05$, $r_p^2 = 0.76$, and the shared variance of the ranked data $r_s^2 = 0.56$. However, as there are two distinct clusters again within the scatter plot, one representing hits in the upper right quadrant, and the other representing false alarms in the lower left quadrant, correlations were calculated separately for these two distinct clusters. The resulting correlations for hits were $r_p(94) = 0.16$, $p = 0.13$, $r_s(94) = 0.09$, $p = 0.40$, $r_p^2 = 0.03$, and the shared variance of the ranked data $r_s^2 = 0.01$, where the resulting correlations for false alarms were $r_p(92) = 0.35$, $p < 0.01$, $r_s(92) = 0.22$, $p = 0.04$, $r_p^2 = 0.12$, and the shared variance of the ranked data $r_s^2 < 0.01$.

As for BEAGLE, when ignoring the distinction between hits and false alarms, the relationship between the model and empirical results was weak, with correlations of with $r_p(188) = 0.31$, $p < 0.05$, $r_s(188) = 0.34$, $p < 0.05$, $r_p^2 = 0.10$, and the shared variance of the ranked data $r_s^2 = 0.12$. Again, calculating the correlations for hits and false alarms separately, the corresponding correlations for hits were $r_p(94) = -0.27$, $p = 0.01$, $r_s(94) = -0.26$, $p = 0.01$, $r_p^2 = 0.07$, and the shared variance of the ranked data $r_s^2 = 0.07$, with the corresponding correlations for false alarms being $r_p(92) = -0.04$, $p = 0.70$, $r_s(92) = 0.01$, $p = 0.90$, $r_p^2 < 0.01$, and the shared variance of the ranked data $r_s^2 < 0.01$.

Results from Experiment 3 provide some evidence that LSA does a better job than BEAGLE at tracking participants' recognition performance. When comparing the two models, all differences were statistically significant in favor of LSA at the $p < 0.05$ significance level, except for the case of the Spearman correlations for false alarms ($z = 1.73$, $p = 0.08$). However, with these statistical differences aside, neither model did a very good job. When evaluating the

relationship between simulated results and empirical results at the item level, LSA did not significantly track hit rates, and BEAGLE did not significantly track false alarm rates. Again, even in the strongest cases of predictions (e.g. high and low false alarm rates for words like *orchestra* and *criminal* in LSA, and *boston* and *crime* in BEAGLE) the empirical results were not consistent with these item specific predictions.

Experiment 4

Experiment 4 was constructed using the same DRM items to derive the distinct study words and new test lures based on BEAGLE rather than LSA. Aside from this difference, the design of Experiment 4 was a replication of Experiment 3. However, due to the COVID-19 pandemic, Experiment 4 was adapted to be run online using jsPsych, a JavaScript library for running behavioral experiments in a web browser (de Leeuw, 2015; https//www.jspsych.org/), and was hosted on GitHub Pages. Data were collected and stored by using Google Firebase (https://firebase.google.com/) before being downloaded and analyzed. The motivation for Experiment 4 was to see if the model predictions from LSA and BEAGLE would differ when the new distinct study words and new test lures were selected when using BEAGLE.

Method

Participants

A total of 105 participants were recruited online via the University of Manitoba SONA subject pool.

Apparatus

Participants access of the experiment was via web browser, where the experiment was presented as a HTML webpage delivered by a combination HTML, CSS, and JavaScript code.

Materials

For the category study words and for the new unstudied words that were related to those categorized words, the same eight lists that were selected for Experiment 3 were used in Experiment 4. The other study and test materials (the studied distinct words, the unstudied words related to the distinct words, and the unstudied unrelated words) were selected in the same fashion as in Experiment 3, however I used BEAGLE rather than LSA vectors trained on the TASA corpus. Furthermore, the pool of 1,996 nouns that was used to select lures from in Experiment 3 was expanded to include the top 5,000 lemmas occurring in the Corpus of Contemporary American English (http://www.wordfrequency.info/). After removing any duplicated words or multi-word exemplars, the resulting pool consisted of 4,076 high frequency words. This was done as BEAGLE was not able to find enough semantically similar (or dissimilar) words to serve as lures in Experiment 4 using cosine similarity measures. Therefore, the pool of possible lures that BEAGLE sampled from was expanded. Selection again was conducted in an iterative process, where the studied distinct words were selected first, with the words related to those distinct words being selected next, and the unrelated words being selected last. A complete list of the materials used in Experiment 4 appear in the Appendix.

Procedure

Upon launching the experiment in a web browser, participants were presented with a screen that displayed the informed consent form. After they had read the consent form and clicked on a button marked “OK”, they were presented with instructions. As this experiment was conducted online, the instructions included an attention check where participants were required to type in a randomly chosen word that was provided in the instructions. If the participant did not type the correct word in the space provided, the instructions were presented again until this condition was satisfied. Exclusion criteria was such that if it took a participant more than five

attempts to type the correct word during the attention check, they were excluded from analysis. Aside from these differences, the Experiment 4 was mechanistically the same as Experiment 3.

Computational Simulations

Simulation results using the LSA and BEAGLE semantic vectors coupled with MINERVA 2 for Experiment 4 are shown in Figure 5.[6] One thousand independent simulations were conducted with *L* set to 0.05 for LSA and *L* set to 0.04 for BEAGLE. For both simulations, τ in MINERVA 2 was set to 7. The first column shows item recognition rates for the empirical data, and item recognition estimates derived from LSA and BEAGLE. The second column takes these item estimates and collapses them into their respective word classes, again for the empirical data and the simulated LSA and BEAGLE results.

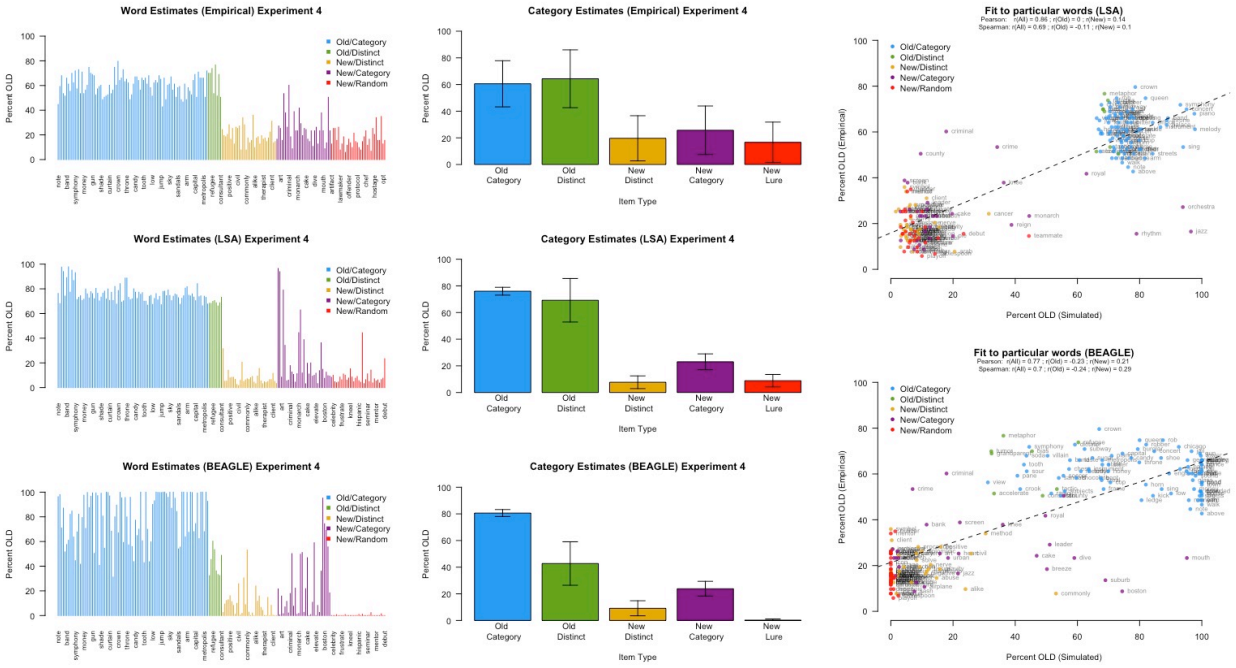

Figure 5. Columns 1 and 2 show item and categorical results for the empirical, LSA and BEAGLE results respectively, in Experiment 4. Error bars represent ± 1 SD. Column 3 shows

[6] The word 'over' was included as one of the category related lures for the *high* list but was omitted from computational simulations as it was not included in the BEAGLE vectors. This oversight was unintentional.

scatter plots with best fitting regression lines for LSA and BEAGLE in comparison to empirical data. Each point represents an individual word.

Firstly, between the different classes of words, both theories predicted that subjects ought to respond "yes" more frequently to old words more than to new words, where old category words had a higher number of "yes" responses than the old distinct words. However, the magnitude for hit rates predicted for the old distinct class of words for BEAGLE was much lower than it was for LSA. Within the new classes of words, both LSA and BEAGLE provided predictions that proposed that the highest proportion of "yes" responses (or false alarms) ought to be for the new category words, followed by the new distinct words, and lastly followed by the new unrelated words. However again, the magnitude for false alarm rates predicted for the new unrelated words for BEAGLE was much lower than it was for LSA.

Although LSA and BEAGLE provided similar categorical predictions, when examining the item estimates, the predictions from both theories were quite different. Both theories predicted that some words ought to be better recognized than others, however there was more variance among the item estimates for BEAGLE than there was for LSA. For both models, there was again more variance among the new classes of words than there was for the old classes of words, particularly for the new category words. For example, within the new category class of words, LSA predicted that the word *jazz* ought to have a very high number of "yes" responses elicited, but a very low number of "yes" responses for the word *county.* As for BEAGLE, within the new category class of words, the theory predicted that participants ought to false alarm to the word *mouth*, but that they ought not to the word *crime*. Therefore, again even in the case of the strongest predictions that LSA and BEAGLE derived, there were differences between the theories in Experiment 4.

Again, given the differences between predictions derived by LSA and BEAGLE, this provided an opportunity to test which model did a better job at tracking participants' recognition behavior, both on the categorical level and item level. Moreover, contrary to Experiment 3, BEAGLE was used to derive the related and unrelated materials to the DRM materials used in Experiment 4. Thus, this again created an opportunity to examine how well BEAGLE did in producing variance among selected items that may be observed among human participants.

Results

Two participants had more than five attempts to correctly respond to the attention check that was presented in the instructions, and their data were excluded from the following analyses. In the results that follow, analyses were conducted on the remaining 103 participants. Empirical category performance of the five different classes of words for Experiment 4 can be seen in the bottom row of Table 1. A repeated measures ANOVA revealed statistically significant differences in "yes" responses as a function of item type, $F(4, 102) = 250.42$, $p < 0.01$, $MSE = 736.90$, $\eta^2 = 0.71$. In terms of the planned comparisons and specific differences between the different classes of words, at the most general level participants responded "yes" to studied words ($M = 60.86$) more frequently than they did to unstudied words ($M = 20.67$), $t(102) = 18.51$, $p < 0.01$. Within the studied words, subjects responded "yes" significantly more frequently to the old distinct words ($M = 64.32$) than to the old categorized words ($M = 60.55$), $t(102) = 2.27$, $p = 0.03$. Between the studied and unstudied words, subjects responded "yes" more frequently to old category words ($M = 60.55$) than to new category words ($M = 25.70$), $t(102) = 16.53$, $p < 0.01$ and responded "yes" more frequently to old distinct words ($M = 64.32$) than to new distinct words ($M = 19.66$), $t(103) = 16.29$, $p < 0.01$. Within the unstudied words, participants responded "yes", or false alarmed more frequently to new category words ($M =$

25.70) than to new distinct words (M = 19.66), $t(103) = 5.77$, $p < 0.01$, and in terms of old category and new category words combined, versus old distinct and new distinct words combined, participants responded "yes" more frequently to the two category classes of words (M 51.25) than the distinct classes of words (M = 28.59), $t(103) = 5.45$, $p < 0.01$. Lastly, the two new classes of words that were related to previously studied words (M = 22.68) (the new category and new distinct words) were better recognized than the new unrelated lures (M = 16.66) that were unrelated to all of the other words, $t(103) = 6.85$, $p < 0.01$.

In sum, the empirical results from Experiment 4 were consistent with the results from Experiment 3 with one exception. Old distinct words had a significantly higher number of "yes" responses than old category words, though this difference was marginal.

Computational Comparisons

Once again, simulation results using the LSA and BEAGLE semantic vectors coupled with MINERVA 2 for Experiment 4 are shown in Figure 5. One thousand independent simulations were conducted with L set to 0.05 for LSA and L set to 0.04 for BEAGLE. For both simulations, τ in MINERVA 2 was set to 7. The first column shows item recognition rates for the empirical data, and item recognition estimates derived from LSA and BEAGLE. The second column takes these item estimates and collapses them into their respective word classes, again for the empirical data and the simulated LSA and BEAGLE results. Lastly, the third column shows the relationship between the empirical results and the predicted results for both LSA and BEAGLE. The scatter plots show a best fitting regression line, with values for Pearson (r_p^2) and Spearman rank-order (r_s^2) correlations. Further correlations are calculated for hits and false alarms separately, as they again appear as two distinct clusters.

On the categorical and item levels, LSA had the same general pattern of results as the empirical data, as seen in columns 1 and 2 in Figure 5, except for the old distinct words. Empirically, old distinct words were better recognized than the old category words. BEAGLE did not have the same pattern of results, as old distinct words were under predicted, as were the new lures. When looking at the particular word recognition rates in column 1, both models had a difficult time predicting item specific hit and false alarm rates compared to the empirical data, especially in the case of BEAGLE.

When comparing the results of LSA and BEAGLE on the item level with the empirical data, we can see the relationships plotted in the third column of Figure 5. The scatter plots here show how well the model predicts that people ought to recognize particular words (on the x-axis) relative to how well they actually recognized those words (empirical data on the y-axis). Ignoring the distinction between old and new words, there was a linear trend for LSA with $r_p(189) = 0.86$, $p < 0.05$, $r_s(189) = 0.70$, $p < 0.05$, $r_p^2 = 0.74$, and the shared variance of the ranked data $r_s^2 = 0.49$. However, as there are two distinct clusters within the scatter plot, one representing hits in the upper right quadrant, and the other representing false alarms in the lower left quadrant, correlations were calculated separately for these two distinct clusters. The resulting correlations for hits were $r_p(94) = 0.04$, $p = 0.70$, $r_s(94) = -0.06$, $p = 0.59$, $r_p^2 < 0.01$, and the shared variance of the ranked data $r_s^2 < 0.01$, where the resulting correlations for false alarms were $r_p(93) = 0.13$, $p = 0.22$, $r_s(93) = 0.31$, $p = 0.76$, $r_p^2 = 0.02$ and the shared variance of the ranked data $r_s^2 = 0.10$.

As for BEAGLE, when ignoring the distinction between old and new words, there was a linear trend, with $r_p(189) = 0.78$, $p < 0.05$, $r_s(189) = 0.70$, $p < 0.05$, $r_p^2 = 0.61$, and the shared variance of the ranked data $r_s^2 = 0.49$. When calculating the correlations for hits and false alarms

separately, the corresponding correlations for hits were $r_p(94) = -0.23$, $p = 0.02$, $r_s(94) = -0.24$, $p = 0.02$, $r_p^2 = 0.05$ and the shared variance of the ranked data $r_s^2 = 0.06$, with the corresponding correlations for false alarms being $r_p(93) = 0.21$, $p = 0.05$, $r_s(93) = 0.29$, $p = 0.01$, $r_p^2 = 0.04$, and the shared variance of the ranked data $r_s^2 = 0.08$.

Taken together, results from Experiment 4 provide some evidence that the models can (at times) track the amount of variance better than random, but both came far from tracking item specific recognition rates. Both models agree that people should be able to recognize studied from unstudied words, but both fail the critical test of predicting item specific recognition rates. Specifically, LSA could not track hit or false alarm rates, and discerningly, BEAGLE inversely predicted the rate of hit rates. When comparing the overall performance of both models, neither model outperformed the other in a statistically significant fashion except for when ignoring the distinction between hits and false alarms, where the overall r_p^2 for LSA captured more of the variance than did the overall r_p^2 for BEAGLE ($z = 2.64$, $p < 0.05$). Finally, even in the strongest cases of predictions (e.g. high and low false alarm rates for words like *jazz* and *county* in LSA, and *mouth* and *crime* in BEAGLE) the empirical results were not consistent with these item specific predictions.

Experiment 5

The overall goal of this thesis was to model people's recognition behavior at the general level in Experiments 1-4, and at the item level in Experiments 1, 3 and 4. However, those experiments were designed with particular stimuli in mind. Thus, the goal for Experiment 5 was to predict recognition behavior by formulating predictions with MINERVA 2 a priori without complications of matching words to categorized lists and themes. Doing so would provide a true test of the extent to which the model can match people's recognition, when there is no ancillary

experimental manipulations on the materials. To test the theory in the strongest case, words were picked at random to test if the theory could predict word specific recognition rates based on semantic relationships derived from LSA and BEAGLE. To implement the study words, they were sampled at random from the TASA corpus. Lists were also of equal length to those used in Experiments 1-4, and dimensionality of the semantic vectors were kept constant. These semantic vectors were then imported to MINERVA 2 as before, and a simulation of recognition performance was conducted. The materials in this simulation were then used to conduct a recognition experiment with the same procedure as in Experiment 4, and results were compared to the predictions from the simulations. The extent to which the model predictions matched people's recognition provides a hard test of the extent to which the models can match people's recognition behavior. Again, comparison between experimental results and model results was tested by computing the resulting R^2 by means of linear regression.

Participants

In the same fashion as in Experiment 4, 98 participants were recruited online from the University of Manitoba subject pool. This sample size again, was large enough to measure item specific recognition rates.

Apparatus

Experiment 5 was conducted online. The apparatus was identical to that in Experiment 4.

Materials

Consistent with Experiments 1-4, 192 words were randomly sampled, without replacement, from the TASA corpus. Half of the words served as study items; the other half served as new lures. A complete list of the stimuli used in Experiment 5 appears in the Appendix.

Procedure

The procedure was identical to that in Experiment 4.

Computational Simulations

Simulation results using the LSA and BEAGLE semantic vectors coupled with MINERVA 2 for Experiment 5 are shown in Figure 6. One thousand independent simulations were conducted with *L* set to 0.05 for LSA and *L* set to 0.007 for BEAGLE. For both simulations, τ was set to 7. The first column shows item recognition rates for the empirical data, and item recognition estimates derived from LSA and BEAGLE. The second column takes these item estimates and collapses them into their respective word classes (in Experiment 5, only old and new words), again for the empirical data and the simulated LSA and BEAGLE results.

The simulation results from LSA and BEAGLE did not provide the same predictions for Experiment 5. As compared to BEAGLE, LSA predicted a higher number of "yes" responses for the old studied words, but a lower number of "yes" responses for the new unstudied words. Furthermore, BEAGLE predicted quite a bit more variance among items than LSA did, with both theories predicting more variance among the new words than the old studied words. For example, within the new words, LSA predicted that the word *unimproved* ought to have a very high number of "yes" responses elicited, but a very low number of "yes" responses for the word *exaggerated*. As for BEAGLE, within the new category class of words, the theory predicted that participants ought to false alarm to the word *fished*, but that they ought not to the word *squalled*. Therefore, even in the strongest cases of predictions that LSA and BEAGLE derived, there were differences between the theories.

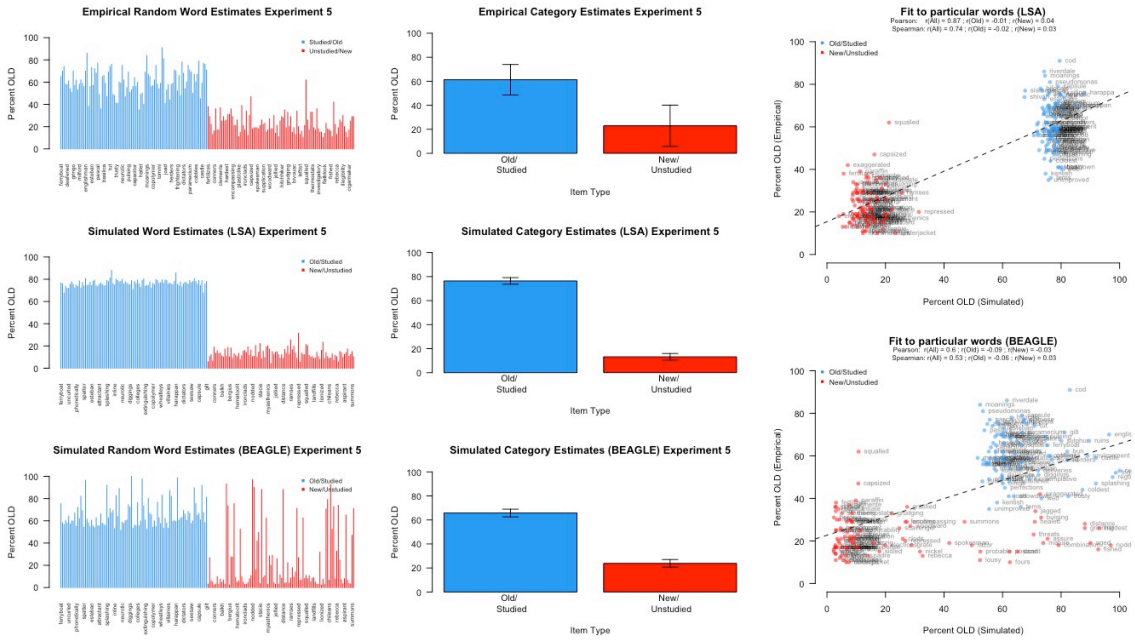

Figure 6. Columns 1 and 2 show item and categorical results for the empirical, LSA and BEAGLE results respectively, in Experiment 4. Error bars represent ± 1 SD. Column 3 shows scatter plots with best fitting regression lines for LSA and BEAGLE in comparison to empirical data. Each point represents an individual word.

Results

Three participants were again excluded for exceeding more than 5 attention checks presented in the instructions. Thus, the analyses that followed were conducted on that data from the remaining 95 participants. The expectation a priori was that the theory would track recognition performance.

Participants responded "yes" to old targets more frequently than to new unstudied lures, $t(94) = 17.50$, $p < 0.01$. Unlike the previous 4 experiments, no further contrasts were conducted on the empirical data as Experiment 5 was primarily conducted to compare the performance of participants versus the model at the item level, without a study and test list composed of five different kinds of items.

Computational Comparisons

Simulation results using the LSA and BEAGLE semantic vectors coupled with MINERVA 2 for Experiment 5 are shown in Figure 6. One thousand independent simulations were conducted with *L* set to 0.05 for LSA and *L* set to 0.007 for BEAGLE. For both simulations, τ was set to 7. The first column shows item recognition rates for the empirical data, and item recognition estimates derived from LSA and BEAGLE. The second column takes these item estimates and collapses them into their respective word classes (in Experiment 5, only old and new words), again for the empirical data and the simulated LSA and BEAGLE results. Lastly, the third column shows the relationship between the empirical results and the predicted results for both LSA and BEAGLE. The scatter plots show a best fitting regression line, with values for Pearson (r_p^2) and Spearman rank-order (r_s^2) correlations. Further correlations are calculated for hits and false alarms separately, as they again appear as two distinct clusters.

As there were only old and new categories in Experiment 5, the category data are presented as such. In the first column of Figure 6, empirical and simulated item results for LSA and BEAGLE are plotted. In the second column, these item estimates are collapsed into old and new classes of words. When comparing the results of LSA and BEAGLE on the item level with the empirical data, we can see the relationships plotted in the third column of Figure 6. The scatter plots here show how well the models predict that people ought to recognize particular words (on the x-axis) based on how they actually recognized words (empirical data on the y-axis). Ignoring the distinction between old and new items, there was a linear trend for LSA with a correlation of $r_p(190) = 0.87$, $p < 0.05$, $r_s(190) = 0.74$, $p < 0.05$, $r_p^2 = 0.76$, and the shared variance of the ranked data $r_s^2 = 0.55$. However, as there are two distinct clusters within the scatter plot, one representing hits in the upper right quadrant, and the other representing false alarms in the lower left quadrant, correlations were calculated separately for these two distinct

clusters. The resulting correlations for hits were $r_p(94) = -0.01$, $p = 0.95$, $r_s(94) = -0.02$, $p = 0.87$, $r_p^2 < 0.01$, and the shared variance of the ranked data $r_s^2 < 0.01$, where the resulting correlations for false alarms were $r_p(94) = 0.04$, $p = 0.70$, $r_s(94) = -0.01$, $p = 0.91$, $r_p^2 < 0.01$, and the shared variance of the ranked data $r_s^2 < 0.01$.

As for BEAGLE, there was also a linear trend when ignoring the distinction between old and new words, with $r_p(190) = 0.60$, $p < 0.05$, $r_s(190) = 0.53$, $p < 0.05$, $r_p^2 = 0.36$, and the shared variance of the ranked data $r_s^2 = 0.28$. Again, calculating the correlations for hits and false alarms separately, the corresponding correlations for hits were $r_p(94) = -0.09$, $p = 0.37$, $r_s(94) = -0.05$, $p = 0.60$, $r_p^2 = 0.01$, and the shared variance of the ranked data $r_s^2 < 0.01$, with the corresponding correlations for false alarms being $r_p(94) = -0.03$, $p = 0.75$, $r_s(94) = 0.03$, $p = 0.74$, $r_p^2 < 0.01$, and the shared variance of the ranked data and $r_s^2 < 0.01$.

The results from Experiment 5 demonstrate that both LSA and BEAGLE agree that people should recognize old words better than new words, but when tracking beyond those two classes, both models had zero predictive power where both did no better than chance at predicting item specific performance. When ignoring the distinction between old and new words, LSA was able to capture more variance than BEAGLE for both the Pearson ($z = 6.22$, $p < 0.05$) and Spearman ($z = 3.50$, $p < 0.05$) correlations, however neither model differed significantly when tracking hits or false alarms separately. Disappointingly, both models did not track item estimates as closely as I had hoped when words were picked at random. Even for the strongest predictions that both models provided (e.g. high and low false alarm rates for words like *unimproved* and *exaggerated* in LSA, and high and low false alarm rates for words like *fished* and *squalled* in BEAGLE), the predictions were not consistent with the empirical results.

General Discussion

The series of experiments in this thesis have shown that an instance theory of recognition memory (MINERVA 2) equipped with semantic vectors from theories of distributional semantics (LSA and BEAGLE) has some degree of validity for predicting word recognition at the general (i.e. category) level but largely and often fails to predict recognition at the word level. Computational theories provide formal and quantitative definitions at precision, where predictions are in turn made at the same level of precision. Having such precise definitions and predictions allows for a theory to be falsifiable, and in the case of this thesis, both LSA and BEAGLE coupled with MINERVA 2 often failed to predict recognition when examining performance at the level of individual words.

Experiment 1 provided an initial test of the theory where stimuli from common categorized word lists were presented in a blocked fashion such that it emphasized the distinctiveness of the distinctive/critical words. At the category level, both LSA and BEAGLE coupled with MINERVA 2 did an acceptable job at predicting people's word recognition behavior, however performance declined when predicting item specific performance. Experiment 2 was designed to test the theory further and provided empirical evidence that the blocked presentation of category words was not an integral part of the paradigm to generally observe the same pattern of results. Experiment 3 expanded upon this paradigm into the realm of DRM materials, utilizing stimuli from Stadler et al., (1999). Moreover, instead of deriving study and test materials (the distinct words, new distinct lures related to those distinct words, and new unrelated lures) in an intuitive fashion as in Experiment 1 and 2, LSA was used to systematically select these experimental materials in Experiment 3. This was accomplished by using cosine measurements of distance within the semantic space — to select words that were semantically similar or dissimilar. Words that were closer in the semantic space were deemed to be more

similar, and words that were further away, were deemed to be more orthogonal. Experiment 4 was mechanistically the same as Experiment 3, where BEAGLE was used to select the aforementioned study and test materials, instead of LSA. Finally, Experiment 5 was constructed to formulate recognition predictions with MINERVA 2 when words were selected entirely at random.

Empirically in Experiments 1-3, the same general pattern of recognition rates for the five different classes of words emerged. Within the old studied words, old category words were better recognized than old distinct words. Within the new or unstudied words, new category words had the highest rate of false alarms, followed by the new distinct words, with the fewest number of false alarms emerging for the new lures. The one exception for this pattern was in Experiment 4, where the old distinct words were better recognized than the old category words, though the magnitude of this difference was not statistically significant.

According to dual process models of recognition and fuzzy trace theory (FTT), it is assumed that recognition judgements are based on two different independent processes. The first is a fast and frugal process operating on familiarity processes known as the gist representation, and the second being a slower but more accurate process, known as a verbatim representation (Reyna et al., 2016; Rotello et al., 2000). Retrieval of a gist trace is where memory for the meaning of items is thought to be represented, where memory for surface features, such as the perceptual details of a stimulus are thought to be represented in a verbatim trace. From these two representations often emerges a recall-to-reject process, where false alarms to associated distractors (e.g. category related words or distinct related words) emerges due to the retrieval of a gist trace, and where correct rejection of an associated distractor emerges from the retrieval of a verbatim trace. FTT is also often used to explain the false memories that emerge from the DRM

paradigm (Soro, et al., 2017). Both of these processes could be used to explain the pattern of results in Experiments 1-3, where participants were more confident in rejecting distinct related lures, utilizing verbatim traces, as they only saw one distinct word that was related during the study phase, where participants were presented with many more category related words associated with the new category lures during encoding, thus consequently utilizing gist traces in memory. As for Experiment 4, it is possible that BEAGLE chose words that emphasized the distinctiveness of the selected critical words, so much so that it initiated the recruitment of a gist trace over a verbatim trace.

Moreover, in Experiment 2 only, the higher number of "yes" responses for the old category words over the old distinct words was statistically significant. This would constitute as a non-von Restorff outcome. A possible explanation for this is that the presentation of the critical distinctive words was not such that it emphasized the distinctiveness of those items. Given that the presentation of all items in the study phase in Experiment 2 and onward were completely randomized, these distinctive words could have appeared in any position (contrary to a typical von Restorff design). Thus, it is possible that the spreading activation among the larger amount of category words created a scenario where these items were easier to recognize than the fewer distinct words.

Conversely, to explain the elevated false alarm rates for the category related lures, another theory posits that there may be more converging activation on these items. The activation-monitoring framework posits that memory is organized conceptually, where the activation of a particular studied word will spread to associated/related concepts or words, which can operate in a cumulative fashion as more related words are studied, which in turn can affect

retrieval processes, where the amount of converging activation on related lures become hard to reject (Soro, et al., 2017).

Whereas MINERVA 2 does not account for FTT or the activation-monitoring framework, it is a result that falls out of MINERVA's familiarity-based process that operates at retrieval. The fact that these mechanisms do not need to be specified within the model can be considered a strength of the model, which can in part account for MINERVA's breadth of success in a number of different memory phenomena.

Where prototypical category exemplars from common taxonomic categories were used in Experiments 1 and 2, DRM lists were used in Experiments 3 and 4. Specifically, the lists of words in Experiments 1 and 2 consisted of words that had explicit prototypical category membership, whereas the DRM lists that were used were constructed of semantically related words. A potential benefit of using both types of categorized lists was that I was able to test if the theory was able to account for both types of semantic relationships.

Admittedly the predictions and model fits formulated by LSA and BEAGLE were not as precise as I had hoped. Both models agree that old words ought to be better recognized than new words, but when looking at item specific prediction rates (tracking hits and false alarms separately), both models fail to do an inadequate job.

It is possible that differences in distinctive word, related lure, and unrelated lure selection may have been responsible for these failures. In Experiment 1 these materials were selected in an intuitive fashion, in Experiment 3 LSA was used, and in Experiment 4 BEAGLE was used to derive these materials. Empirically, participants' recognition performance differed between the five different classes of words, but both models were not sensitive enough to be able to capture these differences. However, the accumulation of more experimental evidence is needed to

confirm or refute this claim. Moreover, other attempts by other researchers to attempt to account for item specific recognition performance have fallen short as well (Osth et al., 2019). In previous literature, most often only categorical performance for classes of words has been accounted for (e.g. Shiffrin et al., 1995).

Moreover, at the savings of computational efficiency, it is possible that not including the order vectors when obtaining the BEAGLE representations could have led to a cost in terms of the selection of words to construct the lures in Experiment 4, and overall performance for the simulations in this thesis. Order vectors have a strong respect for grammatical class among related words, due to the constraints of positional learning. In Osth et al., (2019) they point out that a word like "perform" would become more similar to other words related to artistic expression, such as "write" or "sing", and where a word like boat would become not just similar to other water-related words, but similar to other vehicles as well when including the order vectors in BEAGLE. Indeed, including order vectors would be a simple next step in this line of research.

Certainly, my approach is not the first to examine the issue of distinctiveness using computational approaches. Although I have chosen to examine semantic distinctiveness by utilizing vector space models of semantics coupled with a classic model of memory, this is by no means the only way to examine distinctiveness in memory. There have been a number of other successful approaches that employ computational models, such as Murdock's D scale (1960a) and Brown, Neath and Chater's SIMPLE model (2007).

Murdock's D scale was one of the first models proposed to measure distinctiveness specifically. In this approach, the D scale tracks serial-position effects of absolute judgement, and are accounted for by transforming scale values by a logarithmic transformation, consistent

with the Weber-Fechner Law. From the study of psychophysics, we have known since the 1800's that the relationship between perception and change of a stimulus for many of the body's senses follows a logarithmic function (Fechner, 1965; Jones, 1974). The result is that each item had a measure of distinctiveness which is a function of the sum of the differences in scale value between that particular item and all other items in the list. Years later, this model would inspire the development of the SIMPLE (scale-independent memory, perception, and learning) model (Brown, Neath and Chater, 2007).

SIMPLE also uses Weberian compression, where the authors offer a telephone analogy, where if looking down a road, the closer the telephone poles are to where one is standing, the further apart they appear. However, when looking at telephone poles further down the road, the closer together they start to appear. The importance of this is that it creates an accurate psychological representation, wherein this is the effect that time is thought to have on memory. Weberian compression works to produce this effect, as when applying a logarithmic function to any set of data, it compresses larger values, or temporally distant items more than it does temporally near items (things that happened more recently). One of the main strengths of SIMPLE is that it accounts for forgetting without decay; it provides a natural explanation as to why the passage of time is associated with forgetting. However, it is important to note that it is not a process model, cannot account for rehearsal, and representation can be considered too minimal.

The argument can be made that models such as SIMPLE can work at a decent level of precision, but its core operations are in the abstract; outside of temporal distinctiveness, the definition of distinctiveness is unclear. What we want is a theoretical and empirical approach that

can get us some distance in the realms of precision, breadth, and in developing a definition in regards to what semantic distinctiveness is.

Therefore, in addition to establishing a formal theory of semantic distinctiveness that accounts for both breadth and precision, my goal was to present a theory that accounts for both process and representation, and that is both prescriptive and descriptive. Meaning is defined by use, and we know that the way language is used can be affected by much more than strict grammatical rules, such as cultural and socioeconomic factors (Ghomeshi, 2010). The model(s) employed within this thesis have language use at the crux of their operation and as such, can be considered a strength of the current approach.

A minor limitation was due to the COVID-19 pandemic, as Experiments 4 and 5 had to be adapted to be conducted online, whereas Experiments 1-3 were conducted in person. However, there is little reason to believe that this adversely affected comparability of results, as a large number of behavioral phenomena have been validated and replicated online (Crump et al., 2013). Furthermore, jsPsych in particular has been validated as a reliable tool for conducting behavioral research online, where the results that have been obtained do not seem to be systematically or reliability different than those obtained in traditional lab settings (Hilbig, 2016).

In the work presented, when looking at the fit to particular words in the scatter plots in the figures for Experiments 1, 3, 4 and 5, we can see that there are two distinct clusters: one for hits and one for false alarms. Future directions to find lists of words that maximize the variance across these different classes of words could potentially be a way to remedy the model fits and increase the accuracy of item estimates that were presented in this thesis. Constructing a pool of words that consists of many different categorized lists to draw words from, would be a simple

extension of Experiment 5, where words would be chosen at random, but constrained to a predetermined set of materials. If the model could formulate a priori predictions that maximize the amount of variance between hits and false alarms, this would ideally lead to more precise item estimates and model fits to the experimental data.

Further, it may be beneficial to construct LSA and BEAGLE vectors trained on other bodies of text than TASA. Modern techniques to scrape the internet from websites such as Wikipedia, current news sources, or even Twitter, offer convenient methods to derive current lexical databases to train vector space models on. An advantage of doing so would be to build semantic vectors that are built upon corpora that are familiar to and that people are exposed to everyday.

The empirical research conducted here has a number of different implications for explaining regularities in human memory. As in DeSoto and Roediger (2014), they warn against the increased occurrence of false memories when there is a close match between probes at test and already studied information. Stated differently, "when people try to judge recognition lures that are quite similar to items that they did experience, they are especially likely to be fooled" (DeSoto & Roediger, 2014). Further, they warn that one should be cautious about relying on confidence of recognition performance when required to make recognition decisions among highly similar events.

However, my main interest was how distinctiveness can be mediated when semantic relatedness is considered. From empirical wisdom, it is likely fair to say that knowledge comes from experience. Both parts of my theory here, the semantic derivation of word vectors and the representation of them in MINERVA 2, are both models that learn from experience, consider

use, and in the latter case, are instance based. My goal was to extend the success of these theories with an account of semantic distinctiveness in a word recognition paradigm.

In an era of file drawer effects, the series of experiments presented in this thesis provided evidence against well-defined predictions. With a bias to only report significant results and positive outcomes, null results such as those found in this thesis often go unreported. In the spirit of scientific progress, null results are just as important as significant ones, as science is an incremental process filled with successes and failures, all of which get us closer to some universal truth. With this comes scientific integrity within this thesis, where not only successes are reported, but many failures as well.

Utilizing computational modelling is able to bring this issue under scientific rigor, where the model committed to precise predictions in a well-articulated fashion. Therefore, this approach allowed my theory to be falsifiable, but also allowed for the production of novel predictions and insights that otherwise may not be possible. This work fits into the larger framework of research that has already been conducted, providing novel insights and serves as a contribution to the existing literature, even though a number of null results were obtained.

Lastly, developing a formal set of theoretical tools to arrive at a precise account may benefit machine learning methods and artificial intelligence tools that behave more like people, and even anticipate thought. Coupling the theories used in this thesis together have provided evidence that gives insights on how distinctiveness in word recognition may work, but not at the item specific level. Furthering our understanding of distinctiveness is crucial to furthering our comprehension of human memory. If we can understand distinctiveness, we will be closer to having a complete theory of memory and cognition.

References

Arndt, J., & Hirshman, E. (1998). True and false recognition in MINERVA2: Explanations from a global matching perspective. *Journal of Memory and Language*, *39*(3), 371-391.

Bodner, G. E., Jamieson, R. K., Cormack, D. T., McDonald, D. L., & Bernstein, D. M. (2016). The production effect in recognition memory: Weakening strength can strengthen distinctiveness. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *70*(2), 93.

Brown, G. D., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological review*, *114*(3), 539.

Burgess, C. (1998). From simple associations to the building blocks of language: Modeling meaning in memory with the HAL model. *Behavior Research Methods, Instruments, & Computers*, *30*(2), 188-198.

Calkins, M. W. (1894). Association. *Psychological Review, 1*, 476-483

Chubala, C. M., Johns, B. T., Jamieson, R. K., & Mewhort, D. J. K. (2016). Applying an exemplar model to an implicit rule-learning task: Implicit learning of semantic structure. *Quarterly Journal of Experimental Psychology, 69*, 1049-1055.

Clark, S. E. (1997). A familiarity-based account of confidence–accuracy inversions in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(1), 232.

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

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, *33*(4), 497-505.

Cook, M. (2018). *The mathematics of clinical diagnosis: cognitively-inspired computational psychiatry.* (Unpublished master's thesis). University of Manitoba, Winnipeg, Manitoba.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS one, 8* (3).

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods, 47*(1), 1-12.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology, 58,* 17–22.

DeSoto, K. A., & Roediger III, H. L. (2014). Positive and negative correlations between confidence and accuracy for the same events in recognition of categorized lists. *Psychological Science, 25*(3), 781-788.

Dewhurst, S. A. (2001). Category repetition and false recognition: Effects of instance frequency and category size. *Journal of Memory and Language*, *44*(1), 153-167.

Dewhurst, S. A., & Anderson, S. J. (1999). Effects of exact and category repetition in true and false recognition memory. *Memory & cognition*, *27*(4), 665-673.

Engelkamp, J. (1995). Visual imagery and enactment of actions in memory. *British Journal of Psychology*, *86*(2), 227-240.

Engelkamp, J. (1998). Memory for actions. Hove, England: Psychology Press/Taylor & Francis.

Fechner, G. T. (1965). Element der psychophysik (HS Langfeld, Trans.). *A source book in the history of psychology*, 66-75.

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

Friendly, M., Franklin, P. E., Hoffman, D., & Rubin, D. C. (1982). The Toronto Word Pool: Norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080 words. *Behavior Research Methods & Instrumentation*, *14*(4), 375-399.

Garaizar, P., & Vadillo, M. A. (2014). Accuracy and precision of visual stimulus timing in PsychoPy: No timing errors in standard usage. *PLoS ONE*, *9*(11), e112033.

Gardiner, J. M., & Java, R. I. (1990). Recollective experience in word and nonword recognition. *Memory & Cognition*, *18*(1), 23-30.

Ghomeshi, J. (2010). *Grammar Matters: The Social Significance of How We Use Language.* Winnipeg, MB: Arbeiter Ring Publishing.

Green, R. T. (1956). Surprise as a factor in the von restorff effect. J*ournal of Experimental Psychology, 52,* 340-344.

Günther, F., Dudschig, C., & Kaup, B. (2016). Latent semantic analysis cosines as a cognitive similarity measure: Evidence from priming studies. *The Quarterly Journal of Experimental Psychology*, *69*(4), 626-653.

Hallett, G. (1967). *Wittgenstein's definition of meaning as use.* New York: Fordham University Press.

Hilbig, B. E. (2016). Reaction time effects in lab-versus Web-based research: Experimental evidence. *Behavior Research Methods, 48*(4), 1718-1724.

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

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological review*, *95*(4), 528.

Hunt, R. R. (1995). The subtlety of distinctiveness: What von Restorff really did. *Psychonomic Bulletin & Review*, *2*(1), 105-112.

Hunt, R. R., Worthen, J. B. (Eds.). (2006). *Distinctiveness and memory.* New York, NY: Oxford University Press.

Jamieson, R. K., Mewhort, D. J. K., & Hockley, W. E. (2016). A computational account of the production effect: Still playing twenty questions with nature. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *70*(2), 154.

Jamieson, R. K., & Spear, J. (2014). The offline production effect. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *68*(1), 20.

Johns, B. T., Gruenenfelder, T. M., Pisoni, D. B., & Jones, M. N. (2012). Effects of word frequency, contextual diversity, and semantic distinctiveness on spoken word recognition. *The Journal of the Acoustical Society of America*, *132*(2), EL74-EL80.

Jones, F. N. (1974). History of psychophysics and judgment. *Handbook of perception*, *2*, 1-22.

Johns, B. T., Jones, M. N., & Mewhort, D. J. K. (2019). Using experiential optimization to build lexical representations. *Psychonomic bulletin & review, 26*(1), 103-126.

Jones, M. N., Johns, B. T., & Recchia, G. (2012). The role of semantic diversity in lexical organization. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *66*(2), 115.

Jones, M. N., & Mewhort, D. J. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological review*, *114*(1), 1.

Körding, K. P., & Wolpert, D. M. (2006). Bayesian decision theory in sensorimotor control. *Trends in cognitive sciences*, *10*(7), 319-326.

Kornell, N., & Terrace, H. S. (2007). The generation effect in monkeys. *Psychological Science*, *18*(8), 682-685.

Laham, D. (1997). Latent semantic analysis approaches to categorization. In *Proceedings of the 19th annual conference of the Cognitive Science Society* (p. 979).

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.

Lewandowsky, S. (1993). The rewards and hazards of computer simulations. *Psychological science*, *4*(4), 236-243.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior research methods, instruments, & computers*, *28*(2), 203-208.

MacLeod, C. M., Gopie, N., Hourihan, K. L., Neary, K. R., & Ozubko, J. D. (2010). The production effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(3), 671.

McDaniel, M. A., Waddill, P. J., & Einstein, G. O. (1988). A contextual account of the generation effect: A three-factor theory. *Journal of Memory and Language*, *27*(5), 521-536.

Mikolov, T., Chen, K., Corrado, G., Dean, J., Sutskever, L., & Zweig, G. (2013). word2vec. *URL https://code. google. com/p/word2vec*.

Murdock Jr, B. B. (1960). The distinctiveness of stimuli. *Psychological review*, *67*(1), 16.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, *89*(6), 609.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning* (No. 47).

University of Illinois press.

Osth, A. F., Shabahang, K., Mewhort, D., & Heathcote, A., PhD. (2019, April 2). Global semantic similarity effects in recognition memory: Insights from BEAGLE representations and the diffusion decision model. https://doi.org/10.31234/osf.io/yda2r

Ozubko, J. D., Major, J., & MacLeod, C. M. (2014). Remembered study mode: Support for the distinctiveness account of the production effect. *Memory*, *22*(5), 509-524.

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1-2), 8–13.

Peirce, J. W. (2008). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics*, *2*.

Peirce, J., & MacAskill, M. (2018). *Building experiments in PsychoPy*. Los Angeles, California: Sage.

Pennington, J., Socher, R., & Manning, C. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*(1-2), 73-193.

Recchia, G., Sahlgren, M., Kanerva, P., & Jones, M. N. (2015). Encoding sequential information in semantic space models: Comparing holographic reduced representation and random permutation. *Computational intelligence and neuroscience*, *2015*, 58.

Reyna, V. F., Corbin, J. C., Weldon, R. B., & Brainerd, C. J. (2016). How fuzzy-trace theory predicts true and false memories for words, sentences, and narratives. *Journal of applied research in memory and cognition*, *5*(1), 1-9.

Roediger, H.L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(4), 803–814.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review, 65*(6), 386.

Rotello, C. M., Macmillan, N. A., & Van Tassel, G. (2000). Recall-to-reject in recognition: Evidence from ROC curves. *Journal of Memory and Language, 43*(1), 67-88.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of English verbs. Cambridge, MA: Brandford Books / MIT Press.

Rumelhart, D. E., & McClelland, J. L. (1987). Learning the past tenses of English verbs: Implicit rules or parallel distributed processing. *Mechanisms of language acquisition*, 195-248.

Shiffrin, R. M., Huber, D. E., & Marinelli, K. (1995). Effects of category length and strength on familiarity in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(2), 267.

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of mathematical psychology*, *1*(1), 54-87.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317-1323.

Singer, M., Fazaluddin, A., & Andrew, K. N. (2011). Distinctiveness and repetition in item recognition. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 65*(3), 200.

Simon, H. (1991). *Models of my life.* New York: Basic Books.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of experimental Psychology: Human learning and Memory*, *4*(6), 592.

Soro, J. C., Ferreira, M. B., Semin, G. R., Mata, A., & Carneiro, P. (2017). Ad hoc categories and false memories: Memory illusions for categories created on-the-spot. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(11), 1779.

Steyvers, M., Shiffrin, R. M., & Nelson, D. L. (2005). Word association spaces for predicting semantic similarity effects in episodic memory. In A. Healy (Ed.), *Cognitive psychology and its applications: Festschrift in honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer.* Washington, DC: American Psychological Association.

Surprenant, A. M., & Neath, I. (2009). *Principles of memory.* New York, NY: Psychology Press.

von Restorff, H. (1933). Über die wirkung von bereichsbildungen im spurenfeld. *Psychologische Forschung*, *18*(1), 299-342.

Thomas, R. P., Dougherty, M. R., Sprenger, A. M., & Harbison, J. (2008). Diagnostic hypothesis generation and human judgment. *Psychological review*, *115*(1), 155.

Treisman, A. (1985). Preattentive processing in vision. *Computer vision, graphics, and image processing*, *31*(2), 156-177.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive psychology*, *12*(1), 97-136.

Tversky, A. (1977). Features of similarity. *Psychological review*, *84*(4), 327.

Wallace, W. P. (1965). Review of the historical, empirical, and theoretical status of the Von Restorff phenomenon. *Psychological Bulletin*, *63*(6), 410.

Wilson, M. (1988). MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior research methods, instruments, & computers*, *20*(1), 6-10.

Appendix

Items, Organized by Category and Item Type – Experiment 1

Occupations: salesman, lawyer, professor, courier, teacher, reverend, nurse, engineer, carpenter, janitor, plumber

Critical Word: cobra

Category Related Lures: electrician, accountant, dentist, detective

Critical Word Related Lures: python, anaconda, boa, rattlesnake

Unrelated Lures: thut, redden, propaganda, plank

Countries: France, England, Germany, Russia, Spain, Canada, Italy, Mexico, Japan, Sweden, Bulgaria

Critical Word: limbo

Category Related Lures: China, Brazil, Chile, Kenya

Critical Word Related Lures: ballet, waltz, polka, tango

Unrelated Lures: clamp, causal, spies, occasion

Furniture: chair, table, bed, sofa, desk, lamp, bench, dresser, television, stool, ottoman

Critical Word: sleet

Category Related Lures: buffet, nightstand, recliner, futon

Critical Word Related Lures: blizzard, snow, frosty, ice

Unrelated Lures: patrons, clamming, deluded, reins

Elements: argon, calcium, arsenic, lithium, mercury, potassium, radon, uranium, helium, nitrogen, bromine

Critical Word: yacht

Category Related Lures: carbon, fluorine, hydrogen, plutonium

Critical Word Related Lures: barge, canoe, ship, dinghy

Unrelated Lures: jangle, midway, beaverdam, bang

Sports: soccer, baseball, diving, hockey, tennis, cycling, gymnastics, lacrosse, running, squash, badminton

Critical Word: equity

Category Related Lures: archery, basketball, running, football

Critical Word Related Lures: dividend, liabilities, shareholder, assets

Unrelated Lures: vary, impose, pompous, individual

Grammar: adjective, verb, comma, period, semicolon, paragraph, sentence, plural, syntax, fragment, apostrophe

Critical Word: solar

Category Related Lures: comma, singular, noun, hyphen

Critical Word Related Lures: nuclear, wind, electric, coal

Unrelated Lures: kingdom, deform, kind, summer

Shoes: cleats, heels, mukluks, boots, sneakers, clogs, slippers, loafers, flip-flops, moccasins, flats

Critical Word: pine

Category Related Lures: boots, sandals, sneakers, trainers

Critical Word Related Lures: oak, maple, spruce, birch

Unrelated Lures: repair, sick, fury, exterminate

Organs: liver, brain, lungs, heart, kidney, spleen, intestine, stomach, appendix, bladder, esophagus

Critical Word: vote

Category Related Lures: diaphragm, eyes, lungs, thyroid

Critical Word Related Lures: platform, nominee, poll, campaign

Unrelated Lures: dote, calm exorbitant, singe

Items, Organized by Category and Item Type – Experiment 3

Music: note, sound, piano, sing, radio, band, melody, horn, concert, instrument, symphony

Critical Word: artery

Category Related Lures: jazz, orchestra, art, rhythm

Critical Word Related Lures: blood, chest, heart, vessel

Unrelated Lures: sparkle, protect, eager, confine

Thief: steal, robber, crook, burglar, money, cop, bad, rob, jail, gun, villain

Critical Word: child

Category Related Lures: crime, bank, bandit, criminal

Critical Word Related Lures: development, parent, neglect, infant

Unrelated Lures: puzzle, prayer, patience, south

Window: door, glass, pane, shade, ledge, sill, house, open, curtain, frame, view

Critical Word: election

Category Related Lures: breeze, sash, screen, shutter

Critical Word Related Lures: campaign, cast, vote, republican

Unrelated Lures: white, refuse, absent, great

King: queen, england, crown, prince, george, dictator, palace, throne, chess, rule, subjects

Critical Word: element

Category Related Lures: monarch, royal, leader, reign

Critical Word Related Lures: formula, oxygen, symbol, compound

Unrelated Lures: direct, peace, parcel, stable

Sweet: sour, candy, sugar, bitter, good, taste, tooth, nice, honey, soda, chocolate

Critical Word: equity

Category Related Lures: heart, cake, tart, pie

Critical Word Related Lures: date, location, record, reference

Unrelated Lures: spoke, word, incline, impress

High: low, clouds, up, tall, tower, jump, above, building, noon, cliff, sky

Critical Word: sale

Category Related Lures: over, airplane, dive, elevate

Critical Word Related Lures: contract, display, estate, purchase

Unrelated Lures: idle, devote, single, paint

Foot: shoe, hand, toe, kick, sandals, soccer, yard, walk, ankle, arm, boot

Critical Word: oil

Category Related Lures: inch, sock, knee, mouth

Critical Word Related Lures: gas, olive, valuable, crisis

Unrelated Lures: luster, call, cent, diameter

City: town, crowded, state, capital, streets, subway, country, village, metropolis, big, chicago

Critical Word: plant

Category Related Lures: boston, suburb, county, urban

Critical Word Related Lures: animal, seed, anchor, scatter

Unrelated Lures: operator, present, ocean, several

Items, Organized by Category and Item Type – Experiment 4

Music: note, sound, piano, sing, radio, band, melody, horn, concert, instrument, symphony

Critical Word: tumor

Category Related Lures: jazz, orchestra, art, rhythm

Critical Word Related Lures: cancer, cell, tissue, nerve

Unrelated Lures: artifact, celebrity, sensor, demographic

Thief: steal, robber, crook, burglar, money, cop, bad, rob, jail, gun, villain

Critical Word: bias

Category Related Lures: crime, bank, bandit, criminal

Critical Word Related Lures: positive, negative, battery, feedback

Unrelated Lures: survivor, lawmaker, frustrate, credibility

Window: door, glass, pane, shade, ledge, sill, house, open, curtain, frame, view

Critical Word: refugee

Category Related Lures: breeze, sash, screen, shutter

Critical Word Related Lures: diplomatic, civil, immigrant, arab

Unrelated Lures: playoff, infrastructure, offender, kneel

King: queen, england, crown, prince, george, dictator, palace, throne, chess, rule, subjects

Critical Word: tactic

Category Related Lures: monarch, royal, leader, reign

Critical Word Related Lures: method, procedure, commonly, analysis

Unrelated Lures: depict, icon, threshold, protocol

Sweet: sour, candy, sugar, bitter, good, taste, tooth, nice, honey, soda, chocolate

Critical Word: metaphor

Category Related Lures: heart, cake, tart, pie

Critical Word Related Lures: similarity, description, symbol, alike

Unrelated Lures: hispanic, entitle, teammate, supporter

High: low, clouds, up, tall, tower, jump, above, building, noon, cliff, sky

Critical Word: accelerate

Category Related Lures: over, airplane, dive, elevate

Critical Word Related Lures: magnitude, gravity, particle, attraction

Unrelated Lures: chef, seminar, deploy, tablespoon

Foot: shoe, hand, toe, kick, sandals, soccer, yard, walk, ankle, arm, boot

Critical Word: grandparent

Category Related Lures: inch, sock, knee, mouth

Critical Word Related Lures: therapist, custody, abuse, teenager

Unrelated Lures: documentary, hostage, mentor, online

City: town, crowded, state, capital, streets, subway, country, village, metropolis, big, chicago

Critical Word: consultant

Category Related Lures: boston, suburb, county, urban

Critical Word Related Lures: solve, client, expertise, administrator

Unrelated Lures: scenario, builder, opt, debut

Items, Organized by Studied vs. Unstudied – Experiment 5

Studied/Old: ferryboat, shag, shivah, whooped, deafened, uncurled, minerology, dovecote, gringo, carbine, phonetically, deliveries, milford, bun, statuses, spatter, englishman, riverdale, kentish, thursgood, esteban, hydride, pastors, baloney, pascal, attractant, terns, festering, basilica, laminations, splashing, scrupulous, tut, harappa, contemplative, inline, trusty, showdown, environment, meander, neurotic, anemometers, offical, molali, pulsing, diggings, realized, szabo, capacitor, deadlier, colleges, unimproved, hallet, regularly, web, extinguishing, moanings, sulphur, squalls, exposures, copolymer, caucasoids, amine, mowed, tommi, wheatleys, cod, pseudomonas, joad, kumasi, villainies, coldest, herders, crackdown, skeletal, harappan, frightening, quadrupled, mantelshelf, bifocals, dictators, translation, leniency, ed, paramecium, seesaw, generality, ruins, cobbles, hangers, capsule, perfections, castle, siamese, jackknifing, gilt

Unstudied/New: fertilizes, faculties, centrepiece, billionths, conners, codicil, gayly, inexpedient, clemente, balkh, oscillating, sten, hardest, bulging, bergius, gratified, encompassing, assure, corked, hematocrit, plasticlike, probable, paraffin, admissible, ironclads, feder, mabie, sensor, capsized, nodded, aged, sprig, spokesman, razor, stacie, grinning, supplication, vanishes, postcard, myasthenics, woodward, elections, grate, defensiveness, jellied, clods, byelorussians, applicability, hitchhiker, distance, substantiate, fridley, grudging, tads, ramses, sidled, bivouac, padre, jagged, repressed, leftist, tribulation, fours, enriches, squalled, adamant, recompense, sternberg, thermostats, landfills, rhyming, statuary, investigatory, nickel, lionized, underjacket, flatblock, threats, combination, chileans, fished, lousy, exaggerated, guava, rebecca, mature, scavenger, wieldy, illegibility, aspirant, sand, makeups, cigarmaker, gravitating, summons, healed