University of Alberta

OPTIMIZING HAL PARAMETER SPACE FOR PREDICTING LEXICAL ACCESS AND SEMANTIC DECISION LATENCY.

by

Cyrus Shaoul ©

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of **Master of Science**.

Department of Psychology

Edmonton, Alberta
Spring 2008

Library and
Archives Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-45885-3
Our file Notre référence
ISBN: 978-0-494-45885-3

Abstract

HAL (Hyperspace Analog to Language) is a high-dimensional model of semantic space that uses the global co-occurrence frequency of words in a large corpus of text as the basis for a representation of semantic memory. In the original version of the HAL model, many of its parameters were set without any *a priori* rationale. We took an empirical approach to understanding the influence of the parameters on the measures produced by the HAL model. In particular, we wanted to investigate the power of the HAL model's measures of neighborhood density in predicting reaction times in lexical decision and semantic decision tasks. After exploring HAL's parameter space we found that there are optimal sets of parameters for predicting reaction time from HAL neighborhood density. Importantly, these new parameter sets give us measures of semantic density that predict behavioral measures better than the original HAL parameters.

To Kerstin, Darius and Ilona
for their enormous patience.

Acknowledgements

Thank you, Chris, for your time and patience. Many thanks to TaPor and AICT their support. This work was supported by a National Science and Engineering Research Council (NSERC) grant to Chris Westbury.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Psycholinguistics aims to develop theories of linguistic behavior. One of the key components of language is semantics, the comprehension of the meaning of words. This work investigates a class of models of lexical semantics stemming from HAL (Hyperspace Analog to Language, Lund & Burgess, 1996), a mathematical model of the representation of word meaning. This chapter provides a short summary of the previous research in lexical semantics and mathematical models of word meaning. The second chapter introduces the HAL model and the changes that we have made to it. The third chapter reports on our explorations of the HAL parameter space, and how certain HAL parameter sets can predict behavior in psycholinguistic experiments. The fourth and final chapter contains our conclusions.

1.1 A short history of distributional semantics, Pre-HAL

Distributional semantics aims to use word co-occurrence information to represent word meaning. In essence, substitution regularities are taken to provide information about word meaning. Another way of expressing this concept is that the context surrounding a word conveys important information about its meaning (Harris, 1968). Before there was a formal, mathematical description of this concept, there was philosophical argument for an evidence-based model of word meaning. Wittgenstein (1958) proposed that a word's meaning was defined solely by its usage. He felt that words were similar if they were used in similar ways. In his final work, *Philosophical Investigations*, he addressed the long running philosophical debate about the meaning of words. Wittgenstein proposed that the meaning of a word is not determined by the object that it names (the word "chair" does not always mean a device to sit on). This raises the question: Is it possible to write down the meanings of words? Dictionaries contain definitions, but the use of words is not limited to the usages listed in dictionaries. Wittgenstein made a bold argument: the meaning of a word is completely dependent on its context. His philosophy of meaning did away with word/meaning duality, and opened the door to psycholinguistic theories of meaning that would use contextual information in models of semantic representation.

In the era before modern computing, there were attempts to build many types of semantic models. Osgood, Suci, and Tannenbaum (1957) took an approach that used sets of semantic features to model semantic similarity judgments. After collecting enormous numbers of human judgments of a word's rating on 50 different scales (ex: $wet - dry$, $rough - smooth$), similarity was calculated as a distance between two words in this feature space of 50 dimensions (1 dimension for each scale). The number of scales used was relatively small, but the technique proved to be useful.

Modern distributional semantics uses the same idea of geometric distance in a space, but instead of using a small, hand-made semantic feature space, it uses a corpus-derived context space that has many thousands of dimensions. There has been much debate about the plausibility of large vectors in psychological models. As we will see later, Churchland and Sejnowski (1992) have argued that vector coding is biologically plausible as well as being a very effective way to model many types of psychological processes. The aesthetic and philosophical debates about these ideas are ongoing, but the prowess of high-dimensional models of language is hard to refute.

There are psycholinguists who do not agree with the Wittgensteinian idea of distributional semantics, and they have proposed alternate theories of word meaning. Jackendoff (1983, 2002, 2007) prefers rule systems that allow for semantic categorization. The semantic representations that he proposes are sets of preference rules, applied in a certain order, which he called the Parallel Architecture. This type of representation does not include any statistical information about a word's usage. Rather, Jackendoff sees phonology, syntax and semantics as three parallel systems. Each of the three systems generates representations by formation rules and then all the different types of representations are linked by interface rules. Words must be held in working memory along with their interface rules, which are used to build hierarchical structures. This amount of complexity makes the Parallel Architecture difficult to accept when compared with a more parsimonious model. HAL is an extremely simple model that relies on the complexity of its input. In most cases, large amounts of written text are used to build a model, but the model is a very simple combination of memory for co-occurrence and geometric distance.

One question raised by co-occurrence models concerns the size of the input corpus. Is a 160 million or 300 million (or multi-billion) word corpus a reasonable size to model human performance? The original HAL corpus is indeed very large, but the amount of text used in HAL is large because of the perceptual poverty of text. Compared to the full spectrum of sensory input, text is a poor substitute. Distributional models currently have no way to incorporate sensory-motor representations, and there has been strong evidence that modality-specific systems are involved in the representation and use of conceptual knowledge (e.g Barsalou, Simmons, Barbey, & Wilson, 2003). A response to this criticism is obvious: there are no theoretical barriers to extending the input to distributional models beyond written text. Transcripts of spoken language can be used, once they reach the necessary size. It could be possible to create a rich corpus of linguistic and sensory-motor information and provide this information to HAL. This process has begun. Roy et al. (2006) has recorded all the sights and sounds that were perceived by a child (Roy's own son) from birth to age two, and will analyze this data to better understand how linguistic and non-linguistic input contribute to language acquisition.

In essence, distributional semantics is a way of representing statistical patterns of temporal co-occurrence in any type of sensory input or motor activity. Language is made of of comprehension and production which are based on sensory input and motor control. From this point of view there is a potential connection between language, the body, and the statistical properties of word meaning.

1.2 The Theoretical underpinnings of HAL

A mathematical model of language, like HAL, is by its nature, abstract. A criticism that is often aimed at HAL, LSA, and other mathematical models of language is that they have a weak theoretical position. Unlike other psychological models, such as edge detection in visual processing, HAL does not have a well developed neurophysiological foundation to build on. This situation may leave the reader with an uncomfortable feeling: no obvious justification of the architecture from the "bottom up", and evidence from "top down" experimental data that can sometimes be explained by non-vector models. Is there any reason to expect HAL and other vector models to have any relevance at all? Does the model exist in a theoretical vacuum?

An answer to this question has been put forth by neurocomputational philosopher Paul Churchland. Churchland strongly supports the idea of vector representation and mathematics as a neurologically and psychologically plausible model of psychological processing. In his book "A Neurocomputational Perspective: The

Nature of Mind and the Structure of Science" (Churchland, 1989) he explained the connection between vector models of cognition and neurophysiology. At the same time, he demonstrated that vector-based models can shed light on a variety of psychological processes, including *perceptual recognition* and *explanatory understanding* (Churchland, 1989, p. 197). In this section, I will give a brief summary of Churchland's arguments, and how they support HAL's theoretical underpinnings.

Churchland (1989) bases his theory on the cortical architecture of the human brain. In the cortex, the sensory input layers process information in parallel, with a topographic map that corresponds to the physical world. For example, visual input is mapped by retinal position onto visual space in the visual cortex, and auditory input is mapped by frequency onto the auditory cortex. Connections from these map-like areas of the cortex project into association cortex. The information entering the sensory areas can be very easily represented as a vector of numbers (in the same way that an image can be represented by a vector of numbers, with one number for each pixel in a digital video camera). The input from perceptual layers can then be fed forward to other cortical networks that predict or calculate action. What Churchland asks is: Are there simple mathematical vector operations that can simulate the processing of the input to a biological system that will help us understand its behavior?

To find an answer, he looks at the field of neural network models of cognition (also known as connectionist or Parallel Distributed Processing/PDP architectures). The first step of this type of modeling is to reduce the sensory input to a list of numbers. The PDP models perform perceptual recognition by connecting the input and output units using one or more layers of hidden units. This layer corresponds to the cortical areas connected to the sensory areas. These hidden units are connected and trained using a learning algorithm. After training, they are able to perform operations on the input vectors, operations which show that they can recognize and categorize different inputs, even if they are different from the training data. This means that the connection weights of these internal, hidden units are equivalent to a vector of numbers that represent the activation of a prototype pattern related to the input that the model is trying to recognize. The values of this vector can be also understood as a point in a high-dimensional state-space. As the input to the model changes, the network moves to different positions in that state-space. The calculations done by the network can be framed as matrix multiplication that accomplishes coordinate transformations in state-space.

Based on the success of PDP models, Churchland (1989) extrapolates that vector mathematics may be able to explain many types of cognition. Similar to human cognition, PDP models can process vectors with enormous numbers of elements (just looking at the number of neurons and connections in the brain, the number of elements that the brain can process in parallel is very large, approximately 10^{11} non-sensory neurons). The determinant of processing speed is the number of layers, not the number of elements (the size of the vector). The weights of the connections allow the network to partition input state-space, no matter how many dimensions it has. When there is any type of covariance or cohesion in the structure of the input to the system, the right kind of neural network will learn how to partition up the state-space. Internal to the network, in the hidden layers, are the association vectors that associate input with output. Association vectors need not necessarily be conceived of only as an implementation detail, but may plausibly be implicated directly in cognitive processing. These association vectors are representations of basic prototypes, and coordinate space transformations show that the input may be close or far from the prototype in state space.

Churchland (1989) also links prototypes with the concept of attractors in a high-dimensional state-space. Attractors are a set of states that a system will settle into over time. A neural network may settle into a different state if the initial conditions of input change. However, the network may also enter the same state in response to different (but similar) input states, due to the presence of an attractor. The reason is that more units are involved in the process than just those that are required for mapping the input. If many units in a neural network change on the basis of (for example) some visual input, some of those changes will have nothing to do with the visual form of the object. They will rather relate that form to other aspects of the current state. This ability to generalize across similar inputs by entering the same attractor space amounts to prototype recognition, and endows high dimensional models with explanatory power.

The HAL model is in many ways mathematically equivalent to an artificial neural network model with unsupervised training algorithm. The information that is contained in hidden unit connections in a neural network is instead stored in the global co-occurrence vectors for each word. Burgess and Lund (2000) noted that using HAL with a very small window size produces similar results in word meaning clustering to an Elman Simple Recurrent Network (Elman, 1993). If each word in a language is considered to be a prototype, then the mathematics of calculating a distance in HAL space are not only reasonable to assume, they may by of the same type that underly all types of cognition, according to Churchland. To take this one step further, all language processing may boil down to operations in word space and sentence space. Semantic qualia may be taken to be similar to sensory qualia: a unique set of levels of activation at certain layers in coritcal circuits.

Using Churchland's logic, we can bridge the theoretical gap between neurobiology, psycholinguistics and vector models of language (Churchland, 1989). The philosophical and the experimental are converging towards a statistical, distributed model of cognition, and as more experimental evidence accumulates, the power of these types of models will become evident.

1.3 HAL and its progeny

HAL (Lund & Burgess, 1996; Burgess & Lund, 1997; Burgess, 1998; Burgess, Livesay, & Lund, 1998; Burgess & Lund, 2000) uses word co-occurrence to build a vector space that contains contextual information for every word in the language. A vector space is a geometric representation of data that has an ordered set of numbers associated with each point in the space. Each set of numbers defines the point's location in the space, and is called its vector. Each vector has a dimensionality that is equal to the number of numbers in the vector. HAL space is made up of vectors with one dimension for each word in the language. These HAL vectors are much larger than most vectors used in psychological models. For example, instead of requiring three numbers, x, y and z, as we would use to define a point in the three dimensional space we inhabit, we use N numbers to define a word's position in HAL space, where N is the number of words in the language. In the original HAL work, these word vectors had more than 100,000 dimensions.

The HAL model uses the context of a word's usage to find the neighbours of a word by calculating the distance between all word vectors in this space. This model has been adopted and modified by various researchers since it was proposed in 1996. The following section is a survey of the work done on the HAL model by psychologists and computer scientists since its inception.

There were some psycholinguists who collaborated with Curt Burgess in the early days of the HAL model. Buchanan, Burgess, and Lund (1996) used HAL to model deep dyslexia. They found that words with denser neighborhoods produced more errors in deep dyslexics than words with sparser neighborhoods. Buchanan, Westbury, and Burgess (2001) looked at HAL neighborhood effects on lexical decisions. They found that the HAL neighborhood size was a reliable predictor of lexical decision reaction time. Even after removing the contributions of orthographic variables and imageability, there was significant explanatory power from

HAL neighborhood size.

Siakaluk, Buchanan, and Westbury (2003) investigated the ability of HAL to predict performance in a categorization task. They found that HAL semantic density influenced the decision time on a go/no go task that required participants to classify a word as being animal or non-animal. The influence of density was found to be facilitative, where words with denser semantic neighborhoods were processed faster. Yates, Locker, and Simpson (2003) found a similar facilitatory effect of high-density neighborhoods in a lexical decision task that included pseudohomophone foils.

Many computer scientists have taken this kind of memory model and modified it to solve problems in the field of artificial intelligence. Song and Bruza (2001), Song, Bruza, Huang, and Lau (2003), and Song, Bruza, and Cole (2004) have applied the HAL model to problems of concept learning, inference, and information flow. They were able to use HAL vectors as part of an intelligent software agent that makes "aboutness" judgments such as: the sentence "Welcome to the City of Red Deer, Alberta" has nothing to do with a certain ungulate known as *Cervus elaphus*. They do this by combining the vectors for all the words in the sentence and then comparing it to the vector for the concept in question (in this case, "deer").

During the investigations that are reported in this thesis, other researchers have proposed models that are similar to HAL. We describe three of these very recently reported models here:

Rohde, Gonnerman, and Plaut (2007) created the COALS (Correlated Occurrence Analogue to Lexical Semantic) model. It is identical in design to HAL except in the following respects: it uses a correlation operation for both vector normalization and similarity measures, and it removes closed class words from the model. It also uses SVD (Singular Value Decomposition) to reduce the dimensionality of the co-occurrence matrix. SVD is a factorization technique that can be used to calculate a lower-dimensionality approximation of the original, larger matrix. Rohde et al. (2007) show that HAL performs very well on word similarity tasks such as those in TOEFL exam and other similar tests when SVD is applied to the model.

Bullinaria and Levy (2007) analyzed different influences of excluded closed class words, corpus size, window size and distance metrics. They proposed using an information-theoretic metric, Pointwise Mutual Information (PMI) instead of Euclidean distance, and found that PMI improved the accuracy of their model in their semantic task simulations. PMI is a measure of association that is calculated as the ratio between the probability of two words co-occurring given their joint distribution versus the probability of their co-occurrence given only their individual distributions and assuming independence.

Recently Jones and Mewhort (2007) and Jones, Kintsch, and Mewhort (2006) have built a holographic model of lexical memory that they call BEAGLE (which is an acronym for *bound encoding of the aggregate language environment*). It uses a convolution function as a way to model associative memory (Murdock, 1982). Convolution is a mathematical operation that can be applied to any type of co-occurrence vector to encode it into a memory trace vector. Later, the information can be extracted from the memory trace by calculating the correlation between a probe item and the combined memory trace. In BEAGLE, this function is applied to language in such a way that word order information and global co-occurrence information are simultaneously encoded into each vector. BEAGLE has been able to account for many different types of semantic priming effects when the prime-target pairs are related by both pure semantic relationships and associations (Jones et al., 2006). It has also been used to model sentence completion and semantic categorization (Jones & Mewhort, 2007).

1.4 Our goals

The HAL model has much untapped potential. It is able to predict many different kinds of linguistic behavior, and may have the ability to explain new phenomena. The research that we will present is intended to take the HAL model and understand it better. We will explore HAL's parameter space and find out if there are certain areas in that space that produce more accurate predictions of human behavioral measures than HAL's default parameter set. We will also introduce two new semantic decision tasks, and use HAL to explain the the experimental results we obtain from these tasks.

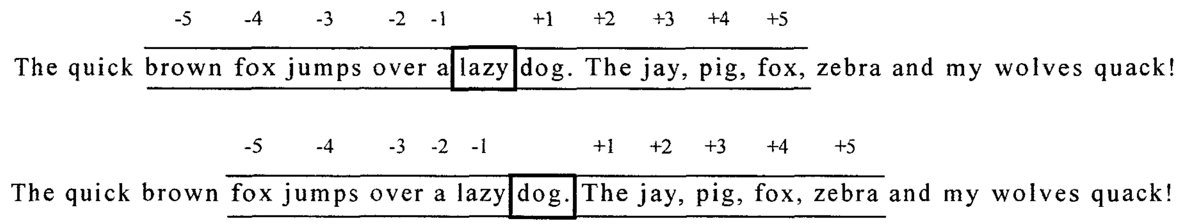
Chapter 2

The HAL model in detail.

The HAL class of models are all based on the original model described by Lund and Burgess (1996). In this chapter I will describe the original HAL model, and then describe HiDEx, a program that implements the HAL model as well as many other very similar models.

2.1 The original HAL model

HAL is a very simple model in many ways. Lexical co-occurrence is captured by keeping track of the number of times all words co-occur with each other within a small window. Words can co-occur when they are adjacent, or when they are separated by other, intervening words. The maximum distance between words considered to co-occur is called the window size. Lexical memories in the HAL model are built by making the model read words in text one window at a time, and then sliding the window forward one word. This process of counting local co-occurrences is illustrated in the Figure 2.1. After reading a whole corpus and counting the local co-occurrences, the data is stored in a raw co-occurrence matrix containing the frequencies of co-occurrence for all possible combinations of words in all possible positions in the window. This matrix can become a very large set of numbers. For example, with a 100,000 word lexicon and a 30 word window, the number of data points in the matrix would be 300 billion.



-5 -4 -3 -2 -1 +1 +2 +3 +4 +5
The quick brown fox jumps over a lazy dog. The jay, pig, fox, zebra and my wolves quack!
-5 -4 -3 -2 -1 +1 +2 +3 +4 +5
The quick brown fox jumps over a lazy dog. The jay, pig, fox, zebra and my wolves quack!

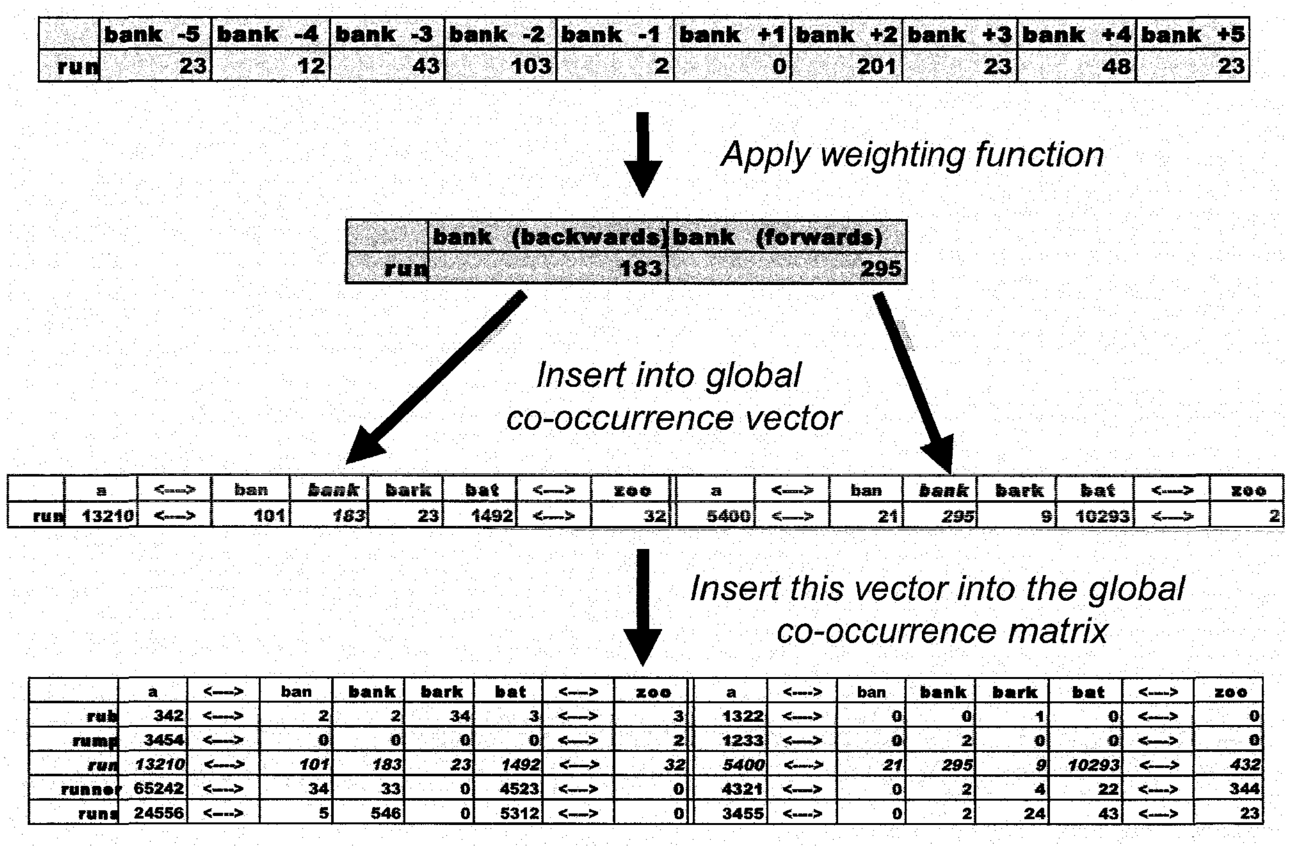
AHEAD	brown	fox	jumps	over	a	lazy	dog	the	jay	pig	zebra
lazy	0	1	0	0	0	0	5	4	3	2	0
dog	0	2	0	0	0	0	0	5	4	3	1

BEHIND	brown	fox	jumps	over	a	lazy	dog	the	jay	pig	zebra
lazy	1	2	3	4	5	0	0	0	0	0	0
dog	0	1	2	3	4	5	0	0	0	0	0

Figure 2.1: A visualization of a sliding 5A5B (five word ahead, five word behind) window as it moves over a sentence. The tables show what the co-occurrence matrix would contain after HAL-style weighting the counts from the sliding window (but before normalizing the rows).

The majority of these data points will contain the number zero, because most words never co-occur with each other. This means that the data in the raw co-occurrence matrix is very *sparse*. To do any meaningful work with the data, it will be condensed or consolidated into a more compact form in the consolidation phase of the HAL model. This consolidation, or aggregation, is done by simply summing the frequencies in the window. There are two parts to the window, the forward part and the backward part, and these two parts are each summed separately. This means that each word will have two numbers for each co-occurrence (See Figure 2.2). The original HAL model used a linear weighting function, called a linear ramp, as a multiplier to give more weight to the words that co-occurred closer to the center of the window. This aggregation reduces the data set from 100 billion to 5 billion data points (by reducing a three dimensional matrix of size 20 x 50,000 x 100,000 to a two dimensional matrix of size 50,000 x 100,000). Each word now has a vector associated with it that contains the aggregate co-occurence for both the forward and backward parts of the window (each vector containing 100,000 numbers)

These vectors are not yet usable due to the the influence of orthographic frequency. Due to the non-linear nature of the distribution of word usage (Zipf, 1935, 1949; R. H. Baayen, 2001), a small number of words will have very high orthographic frequencies, and consequently very high co-occurrence frequencies. The vast majority of the words in a language, on the other hand, will have low frequencies and co-occurrence frequencies. Due to this bias, high frequency words will have vectors that are very dense with large values, and therefore they will be much closer in context space to all words than low frequency words. The original HAL model dealt with this frequency issue by normalizing each vector, by dividing each element in the vector by the vector's length. As we shall see, normalizing vectors in this particular ways leads to a systematic frequency bias, and will not be used in our implementation of the model (Shaoul & Westbury, 2006).



	bank -5	bank -4	bank -3	bank -2	bank -1	bank +1	bank +2	bank +3	bank +4	bank +5
run	23	12	43	103	2	0	201	23	48	23

	bank (backwards)	bank (forwards)
run	183	295

	a	<--->	ban	bank	bark	bat	<--->	zoo	a	<--->	ban	bank	bark	bat	<--->	zoo
run	13210	<--->	101	183	23	1492	<--->	32	5400	<--->	21	295	9	10293	<--->	2

	a	<--->	ban	bank	bark	bat	<--->	zoo	a	<--->	ban	bank	bark	bat	<--->	zoo
rub	342	<--->	2	2	34	3	<--->	3	1322	<--->	0	0	1	0	<--->	0
rump	3454	<--->	0	0	0	0	<--->	2	1233	<--->	0	2	0	0	<--->	0
run	*13210*	<--->	*101*	*183*	*23*	*1492*	<--->	*32*	*5400*	<--->	*21*	*295*	*9*	*10293*	<--->	*432*
runner	65242	<--->	34	33	0	4523	<--->	0	4321	<--->	0	2	4	22	<--->	344
runs	24556	<--->	5	546	0	5312	<--->	0	3455	<--->	0	2	24	43	<--->	23

Figure 2.2: The process of creating a global co-occurrence vector. Local co-occurrence is counted, then weighted, and then inserted in the the global co-occurrence matrix.

The final stage in preparing the vectors for distance calculations is the elimination of the noisy, sparser parts of the matrix. This is done in the original HAL model by only retaining vectors for the words with the greatest row variances. If only the rows with the top 10,000 most variant words are used, the forward and backward aggregates create rows of 20,000 elements. This reduces the size of the final matrix from 50,000 x 100,000 (5 billion elements) to 50,000 x 20,000 (1 billion elements). This matrix is smaller and denser than previous matricies. It is now small enough to fit into the memory of modern computers, making the calculations tractable.

At this point in the process each word in the lexicon has a representation that consists of 20,000 elements (equivalent to 20,000 dimensions). The HAL model uses the Euclidean distance metric to calculate the distance between any two words in the space. For every element j in the vectors for words a and b, $d = \sqrt{\sum_j (a_j - b_j)^2}$. This distance expresses how similar the contexts of usage of the two words are. If the words have similar values in the same dimensions, they will be closer together in the space. To find the neighbors of a word in context space we calculate the distance between the word and all the other words in the language. The closest words are considered as neighbors in HAL space. The neighborhood density is a measure of how tightly packed the words in the neighborhood are. The density measure in the original HAL work was calculated by averaging the distances between the word and its 10 closest neighbors. This produced a density value for each word, conceptually similar to the orthographic neighborhood density, but in a semantic space instead of an orthographic space.

2.2 HiDEx and our modifications to HAL

We have created a novel implementation of the HAL model called HiDEx (High Dimensional Explorer). HiDEx is capable of running the HAL model using the identical calculations that were specified in the work of Lund and Burgess (1996), but it is also able to use slightly different algorithms and parameters. Alternative algorithms include: new normalization algorithms, new weighting algorithms, and new neighborhood membership algorithms. Alternative parameters include new window sizes and context sizes. In this next section we will outline the modifications we made to HAL, and why we made them.

2.2.1 Lexicon choice

The lexicon that we chose to use for HiDEx was derived from the CELEX database (R. H. Baayen, Piepenbrock, & Gulikers, 1995) by choosing all the words that had an orthographic frequency of two occurrences per million or greater. This lexicon contains approximately 45,000 words, which is less that the 70,000 word lexicon used by Lund and Burgess (1996). The choice to reduce the lexicon size was made for two reasons: 1) the amount of information contained in the contexts of low frequency words is small, and does not have much influence on the distances between most words in the space, and 2) the computational complexity of the model increases greatly with the size of the lexicon.

2.2.2 Corpus Choice

Lund and Burgess (1996) used a corpus 160 million words of USENET (Fristrup, 1994) text. It is well known that the balance of registers and genres in a corpus has a strong effect on the HAL vectors produced (Shaoul & Westbury, 2006; Bullinaria & Levy, 2007; Rohde et al., 2007). In order to make our results comparable to the majority of studies done on the HAL model, we chose to replicate as closely as possible the USENET corpora used by Lund and Burgess (1996), Burgess and Lund (1997), Burgess (1998), Burgess and Livesay (1998), Burgess et al. (1998) and Burgess and Lund (2000). We collected 12 billion words of USENET text from 2005 to 2007 (Shaoul & Westbury, 2007), and use a 1 billion word subset of this corpus to build our models. The same benefits that were described by Lund and

Burgess (1996) are true for this corpus: USENET text contains a very broad variety of genres and topics, and most of the text is in a very conversational style, similar in some ways to spoken language. We chose not to use a corpus of 160 million words in size because we found that there were many words in our 50,000 word lexicon that had one or less occurrence in this corpus. To obtain observations of multiple occurrences of all the words in our lexicon, it was necessary to use a larger corpus. Due to the computing time required to run all the experiments in this study, it was impossible to do a comparison of the results based on different sized corpora. In addition, Bullinaria and Levy (2007) did a very thorough analysis of the impact of corpus size on HAL. They found that their measures of performance increased as corpus size increased, but the amount of improvement was mostly at ceiling for corpora of 90 million words or greater. This result leads us to believe that our choice of corpus will not greatly impact our results, and will allow us to compare them to previous work with USENET corpora of smaller size.

2.2.3 Frequency issues/Normalization

Shaoul and Westbury (2006) showed that there was a problem with the original HAL model that allowed a word's orthographic frequency to influence its neighborhood density. If HAL neighborhood density is used to predict psycholinguistic phenomena, it would be unfortunate if HAL density measures covaried with orthographic frequency, one of the most powerful predictors of lexical access (Balota & Spieler, 1999). Shaoul and Westbury (2006) found that the normalization procedure used in the original HAL, dividing each vector by its variance, did not eliminate frequency effects. Buchanan et al. (2001) proposed using the orthographic frequency of each word as the word's vector's divisor, and Shaoul and Westbury (2006) did exactly that. Words with high frequency would see their co-occurrence values shrink, and words with low frequency would see their values amplified. Shaoul and Westbury (2006) found that the neighborhood densities made with this new normalization technique were no longer correlated with orthographic frequency.

2.2.4 Weighting and Window Size

Lund and Burgess (1996) used one method for assigning weights to the co-occurrence counts, the linear ramp, without describing any *a priori* justification for their choice. The original HAL model used 10 word windows, and the values were multiplied by the distance from the end of the window. This meant that the count for the word appearing directly adjacent would be multiplied buy 10, then the next one out by 9, and so on. We introduced 8 alternative weighting functions. For a detailed description of the functions, please see Table (2.1). We also allow the size of the forward and backward windows to be set independently to any size.

Function Name	Function w = window size, p = position (1 to w)	Sample Vector of Weights (symmetric 4-word windows)
Flat Weights	$x = 1$	[1 1 1 1 1 1 1 1]
Linear Ramp	$x = (w - p + 1)$	[1 2 3 4 4 3 2 1]
Exponential Ramp	$x = (w - p + 1)^2$	[1 4 9 16 16 9 4 1]
Forward Linear Ramp, Backward Flat Ramp	$x = 1, x = (w - p + 1)$	[1 1 1 1 4 3 2 1]
Forward Flat Weights, Backward Linear Ramp	$x = (w - p + 1)$,$x = 1$	[1 2 3 4 1 1 1 1]
Inverse Linear Ramp	$x = p$	[4 3 2 1 1 2 3 4]
Inverse Exponential Ramp	$x = p^2$	[16 9 4 1 1 4 9 16]
Second Word Weighting	if $p = 2$, $x = 10$, else $x = 1$	[1 1 10 1 1 10 1 1]
Third Word Weighting	if $p = 3$, $x = 10$, else $x = 1$	[1 10 1 1 1 1 10 1]
Fourth Word Weighting	if $p = 4$, $x = 10$, else $x = 1$	[10 1 1 1 1 1 1 10]

Table 2.1: List of Weighting Functions Implemented in HiDEx.

2.2.5 Context Size

One key part of the HAL model is the reduction of the size of the global co-occurrence matrix after the weighting scheme has been applied and the windows have been summed. The original HAL model sorted all the vectors by variance, and only retained the N vectors with the highest variance.

In the original HAL model N was set to 200, using the 200 most variant word vectors (Lund & Burgess, 1996). HiDEx allows this parameter, which we call *context size*, to be set to any value less than the lexicon size.

2.2.6 Neighborhood size, neighborhood membership threshold

Another extension to HAL proposed by Shaoul and Westbury (2006) was the concept of a neighborhood membership threshold. Unlike HAL, which used a fixed number of the closest neighbors as the neighborhood, we calculated a number, called the membership threshold, that was used as the criterion for neighborhood membership. This threshold is calculated by randomly sampling many millions (usually billions) of word pairs and calculating their inter-word distances to find the standard deviation of this distance distribution. The neighborhood membership threshold was set to: $\mu + 1.5\sigma$ and is used for all future neighborhood calculations. Note that this threshold has to be re-calculated every time any other parameter in the model is changed since the average distance between words will be affected by any parameter change. A consequence of this new definition of neighborhood membership is that some words may have more neighbors than others, and some words may have no neighbors.

2.2.7 Two new measures: NCOUNT and ARC

Shaoul and Westbury (2006) introduced two new measures of semantic density that depend on this threshold. The first, Average Radius of Co-Occurrence (ARC) is calculated by taking the mean of the distances between the word in question and all the neighbors with the threshold. The second, Neighbor Count (NCOUNT) is the number of neighbor words in the threshold. These two new measures both relate information about the density of the context neighborhood of words (See Figure 2.3). In later sections, we will be doing analyses of value called NCOUNT-INV, which is defined as:

$$\frac{1}{NCOUNT + 1} \tag{2.1}$$

This produces a value of one for words with no neighbors, and smaller numbers for words with more neighbors.

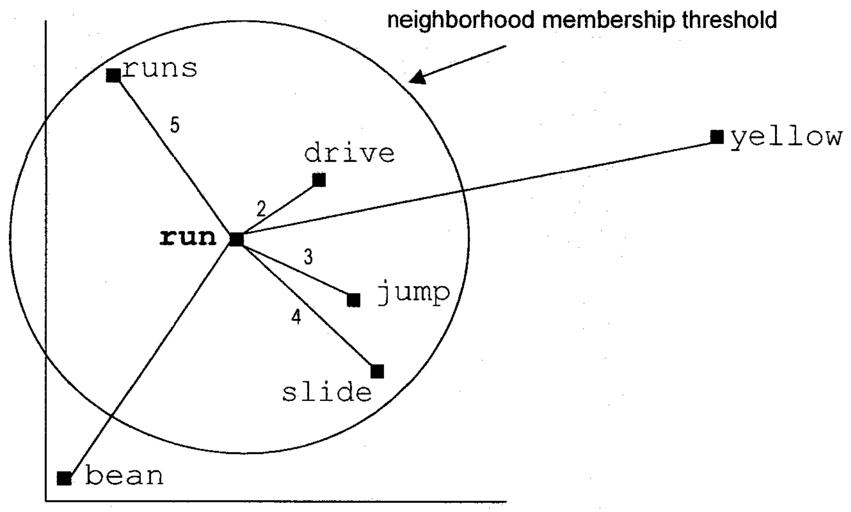


Figure 2.3: A two-dimensional visualization of the neighborhood membership threshold. The word “run” in this example has four neighbours, giving “run” an NCOUNT of 4, and the average distance between “run” and its neighbours, the ARC measure, is 10.25.

2.2.8 Computational Complexity

To give the reader some perspective on the scale of calculations performed by HiDEx, here is a broad, slightly simplified explanation of the process.

- Build Data Set: At the beginning of a set of HiDEx experiments, we must build a Data Set. This step is usually done once, and the Data Set is then re-used for many experiments.
  - Collect documents into a corpus. In our case, we used USENET text as described in Shaoul and Westbury (2007).
  - Initialize with the number zero a matrix that is L x L x N in size, where L is the size of the lexicon, and N is the maximum size of the window that can be used by HiDEx. In the experiments described here $L = 50000$ and $N = 30$, 15 ahead of the center word, and 15 behind. The number of elements in this matrix is 75,000,000,000.
  - Note all word co-occurrences for the center word and each of the other words in the window, and increment the values in the co-occurrence matrix.
  - Slide window forward by one word. Repeat until reaching the end of the corpus.
  - Save this matrix as a Data Set (Approximate Size: 63Gb)
- Calculate Semantic Distances/Densities for a list of words:
  - Load Data Set, and any list of words in the lexicon.
  - Load the desired parameters (window size, weighting scheme, etc)
  - Apply the window size and weighting scheme to the Data Set, consolidating it, creating the global co-occurrence vectors. Retain only the vectors for words with the highest orthographic frequency.

- Normalize the vectors.
- Generate the neighborhood membership threshold by calculating 5% of all the possible pairwise distances (typically on the order of 2 billion distances).
- Calculate distances and neighborhoods for all words in word list.

The process of building a data set currently takes four days of continuous processing by a supercomputer. The process of calculating neighborhoods on a supercomputer takes approximately two hours per parameter set, more if the length of the word list is larger than 500 words. This performance depends on using many CPUs in parallel. HiDEx was designed to take advantage of multi-CPU supercomputers, and has been run using 64 CPUs and 256 gigabytes of memory. The sheer number of calculations required to run these models is enormous, and has deterred many from doing research in this field. The software engineering required is daunting, but we feel that the rewards of this line of inquiry are worth the effort.

Chapter 3

Exploring Parameter Space

The parameters used in the original HAL work were chosen arbitrarily. There was no empirical or theoretical justification given for the choice of window size, weighting function or context size. The questions we wanted to address are: Is there a new set of parameters that will create a better model of word meaning? Will this new parameter set give HAL more explanatory power? Will it shed light on the structure of the mental lexicon? We explored HAL parameter space to find the answers to these questions.

3.1 A coarse-grained exploration of the space

The parameter space of the HAL model that we have implemented is very large; that is, there is a very large number of possible unique combinations of the parameters we have described above (corpus type, corpus size, window sizes, weighting functions, context sizes and others). To evaluate all the possible combinations would take centuries of computation on the supercomputers available to us. We decided to make a preliminary, coarse-grained traversal of the parameter space. We then used a fitness function to find the parameter set that best fit the experimental data. During this exploration, only two parameters were varied at a time while all the other parameters were held constant at their default HAL values. This strategy allowed us to find out how these two parameters influenced the model individually, and how they interacted. Processing time constraints prevented us from exploring three-way or higher interactions, but there is no logical reason to discount their existence. This research is undeniably exploratory in nature, and this coarse-grained approach will allow us to understand how two parameters influence the output of the model.

3.1.1 Parameters that were varied

The two parameters that were varied in this initial exploration of parameter space were window size and weighting function. We chose these two parameters because they are the two parts of the HAL model that have the most potential to change the contextual information stored about words. The size of the window is the only parameter that can change which words are considered to share context. Smaller windows will prevent long-distance contextual relationships from forming. For example consider the following sentence:

> Mathematics, rightly viewed, possesses not only truth, but supreme beauty — a beauty cold and austere, without appeal to any part of our weaker nature, without the gorgeous trappings of painting or music, yet sublimely pure, and capable of a stern perfection such as only the greatest art can show." (Russell, 1910, p.73)

If the window size was only 5 words behind or ahead, there would not be any co-occurrence trace from this sentence for the words "MATHEMATICS" and "BEAUTY".

The weighting scheme that is used to aggregate local co-occurrence across the co-occurrence window has a very slightly different influence on the model's structure. By emphasizing the contextual importance of different parts of the window, it can boost or shrink the influence of proximally co-occurring words. In the above quote, the co-occurrence frequency of "PAINTING" and "MUSIC" could be weighted by a factor of 10 (using the Second Word scheme) or 2 (using the Inverse Ramp scheme). This difference in weighting could significantly change the distance between "PAINTING" and other words in context space, altering the neighborhood and the neighborhood density.

To explore the influence of these two parameters we created a list of all the possible combinations of forward and backward window sizes of zero, five or ten and all the weighting functions listed in Table 2.1. This list contained 73 sets of parameter combinations. In Experiments 1, 2, and 3 we compared the relative predictive power of these 73 parameters sets using two different fitness functions, described in the following section.

3.1.2 Fitness functions

A great virtue of the HAL model is that it is simple and flexible: it will take as input any type of textual material, and produce as output inter-word distances and context neighborhoods. Unfortunately, this flexibility of input makes it very difficult to compare the output of the model when used by different researchers. The relative merits of different model parameters were compared by Bullinaria and Levy (2007). Bullinaria and Levy (2007) used four different methods to compare the fitness of their models: the TOEFL test (using HAL to choose the one word as the correct answer in a multiple choice exam), a Distance Comparison test (comparing interword differences between known semantically related pairs and random pairs), a Semantic Category test (testing if words are closer to the name of their category than to the names of other categories) and a Syntactic Categorization test (testing if a word was closer to its syntactic category center than to other syntactic category centers). These tests provide some information about how the model performs in capturing the structure of the human semantic space. The weakness of these tests is that they depend on handpicked word lists that do not generalize to the rest of the language.

We are more interested in how the HAL model can be used to understand the organization of the mental lexicon, and so we used fitness functions that were based on the correlation between our measurements from the HAL model and behavioral measures of lexical access. The first fitness function was the correlation between a word's ARC, and the average LDRT (lexical decision reaction time) provided by Balota et al. (2002). The lexical decision task has been shown to involve automatic retrieval of semantic information (Balota, Black, & Cheney, 1992). If HAL measures, such as ARC, could explain a heretofore unexplained proportion of the RT variance, then it would validate its ability to model semantic memory. The parameter sets were explored in Experiment 1.

In Experiments 2 and 3, we collected our own SDRT (semantic decision reaction time) data from a semantic decision task. There has been no previous research on how the HAL model can be used to predict semantic decision reaction time. We proposed the following HAL measures as potential candidates that might be predictive of semantic decision reaction time:

- Inter-word distance: This is the distance between any two words in HAL space. This HAL measure may influence reaction time by facilitating the retrieval of lexical semantics due to the priming effects of the retrieval of two words with similar contextual history.

- ARCs and NCOUNTs: These are the ARC and NCOUNT measures for each word in the pair of words. The ARC and NCOUNT measures for the first and second words in a pair could influence reaction time as well. From previous work with LDRT (Buchanan et al., 2001, Shaoul & Westbury, 2006), we saw that words with sparser neighborhoods showed faster lexical decision reaction times (if other lexical factors are all held equal). This could imply that the ARC or NCOUNT of either word could influence the semantic decision reaction times in experiments 2 and 3.

- ARCSUM and NCOUNTSUM: These are the sum of the ARCs and NCOUNTs for the words that make up the pair. The summed ARC and NCOUNT capture the combined densities of the words in the pair.

We used the same 73 parameter sets to calculate the above seven context measures and then tested the strength of the relationships between these measures and the mean SDRT of each item. As this is uncharted territory, the only way to find out if any of these measures are useful in predicting SDRT is to calculate their values using different HAL parameter sets, and compare the relationships using a correlational analysis.

Our prediction is that these measures of contextual similarity, neighborhood density and combined neighborhood size will explain the variability in reaction times for semantic decisions. In the following experiments we will explore the ability of HAL measures to predict lexical decision and semantic decision performance.

3.2 Experiment 1: Predicting Lexical Decisions Reaction Time

3.2.1 Method

For our first experiment, we chose to use the LDRT data from the English Lexicon Project (Balota et al., 2002) as our dependent measure. We obtained averaged LDRT data for 40,481 words (averages for each word across participants) and used this data to run simulations of lexical decision experiments. In each simulation, a random subset of 500 words was sampled from the 40,481 in the list. Then HiDEx was used to calculate the the ARC and NCOUNT-INV for these 500 words using one set of parameters. This process was repeated 73 times, with different random lists of 500 words and all the desired parameter sets. We used the technique of sampling randomly from a large set of words to avoid over-fitting our model.

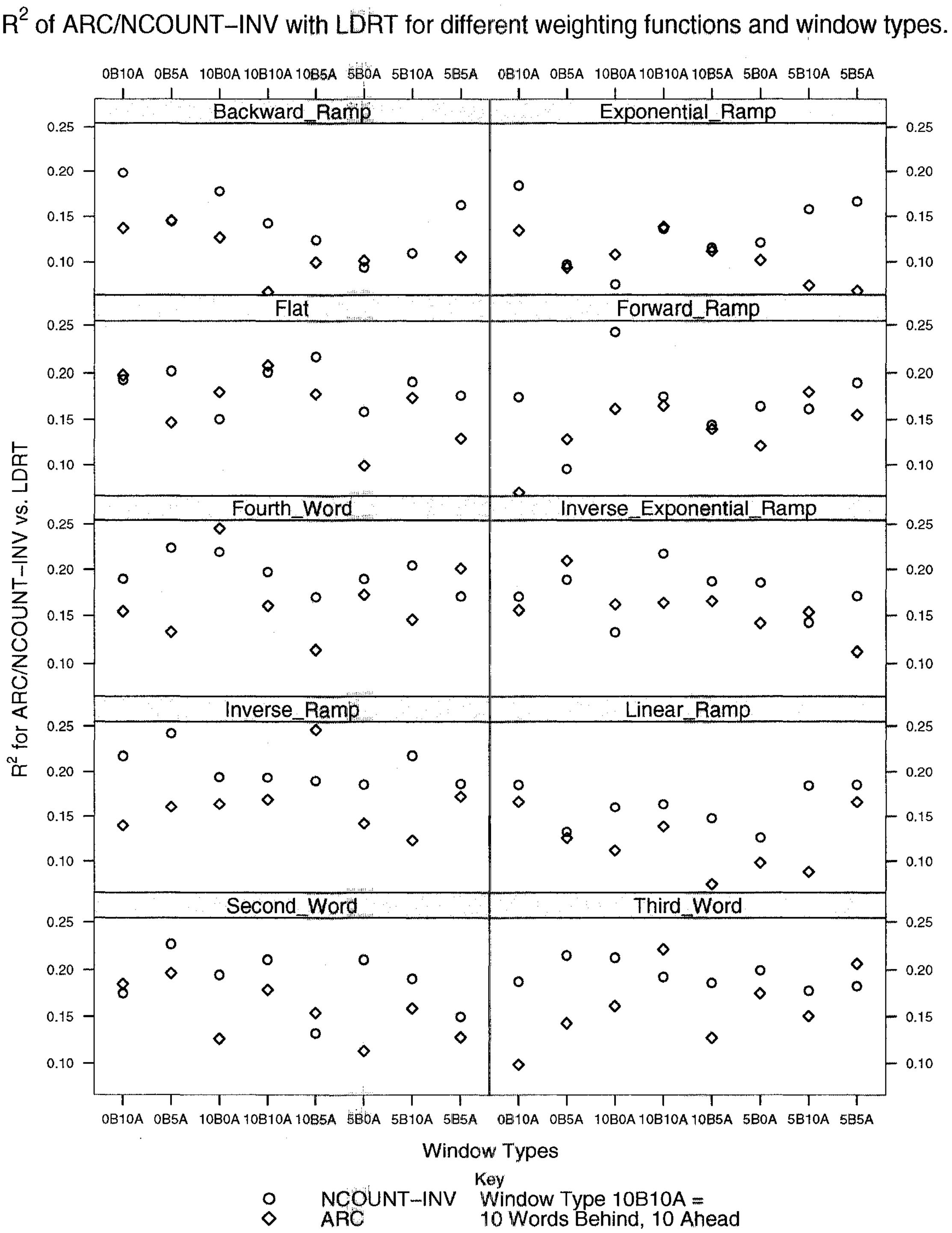


Figure 3.1: Overview of R^2 of NCOUNT-INV and ARC with LDRT for different weighting functions and window types. All correlations are significant ($p < 0.001$)

3.2.2 Results and Discussion

We computed the correlation between LDRT and ARC for each of the 73 parameter sets. There was a large amount of variation in the correlations between LDRT and ARC and NCOUNT-INV for the sets of parameters we tested. The means, ranges and standard deviations of the regressions of ARC were $\mu_{R^2_{ARC}}$ = 0.14, $\sigma_{R^2_{ARC}}$ = 0.041, ranging from 0.04 to 0.25. The means, ranges and standard deviations of the regressions of NCOUNT-INV were $\mu_{R^2_{NCOUNT-INV}}$ = 0.17, $\sigma_{R^2_{NCOUNT-INV}}$ =0.035, ranging from 0.08 to 0.24.

The most important correlations are shown in Figure 3.1. The three window sizes with the highest median R^2 with LDRT were 0B10A[1], 5B10A, and 5B5A, and the weighting functions that performed well with many different window sizes were: Inverse Linear Ramp, Inverse Exponential Ramp, Third Word and Fourth Word. The best combination of parameters for LDRT predicting ARC were: Inverse Exponential, 10B0A (R^2_{ARC} = 0.26). The original HAL parameters (Linear Ramp, 10B10A) produced a much smaller correlation (R^2_{ARC} = 0.11). For NCOUNT-INV, Inverse Exponential, 0B10A was the most predictive ($R^2_{NCOUNT-INV}$ = 0.25). The original HAL parameters again produced a much smaller correlation ($R^2_{NCOUNT-INV}$ = 0.14). Graphs of the parameter sets sorted by median R^2 are shown in Figures 3.2,3.3,3.4, and 3.5.

Since we used a random sample of words in each of these simulations to demonstrate the generalizability of HAL to the whole lexicon, we cannot directly compare the correlations that we have calculated to test for statistical significance. In order to confirm the increase in correlation for our parameter sets, we chose a random subset of 5000 words from the lexicon, and calculated the neighborhoods for these words using two parameter sets of interest: the original HAL parameter set, and one of the best of our 73 parameter sets (Inverse Ramp, 10B5A). The results of this

[1] This notation, 0B10A, is a condensed expression of "No words behind, 10 words ahead"

comparison are shown in Table 3.1. The difference between the explanatory power of the two models can be seen in the size of the AIC (Akaike Information Criterion) value. The AIC is calculated using the following equation:

$$AIC = 2k - 2\text{log-likelihood} \tag{3.1}$$

where k is the number of parameters and log-likelihood is the natural logarithm of the likelihood function of the model in question. There is a difference of over 100 between the AIC scores for both the ARC and NCOUNT-INV models , meaning that we can select the optimal models over the original models because they are much more likely given the data.[2]

Parameter Set	**Weighting Function, Window Size**	R^2_{ARC}	AIC (ARC)	$R^2_{NCOUNT-INV}$	AIC (NCOUNT)
Original HAL	Linear Ramp, 10B10A	0.12	61660	0.17	61349
Optimized HAL	Inverse Ramp, 10B5A	0.15	61440	0.18	61231

Table 3.1: Comparison of correlations between LDRT and ARC/NCOUNT-INV for the original HAL parameters and an optimized set of parameters

Are these relationships stable? To avoid any contamination of the results by spurious correlations, and to validate the results we obtained, we needed to be certain that our results were stable across different subsets of the lexicon. We measured the stability of these correlations across two different sets of words. The same sets of parameters were re-run with different random sets of 500 words, and the average absolute difference in R^2 between runs was 0.04 (LDRT-ARC) and 0.03 (LDRT-NCOUNT-INV). This small amount of difference between runs shows that the correlations are stable across different random samples of words.

[2] We also calculated the BIC (Bayseian Information Criterion) for these models, and saw the same result: the optimized models had smaller BIC values. For ARCs, the BICs were 61680 and 61460 and for NCOUNT-INV the BICs were 61369 and 61251.

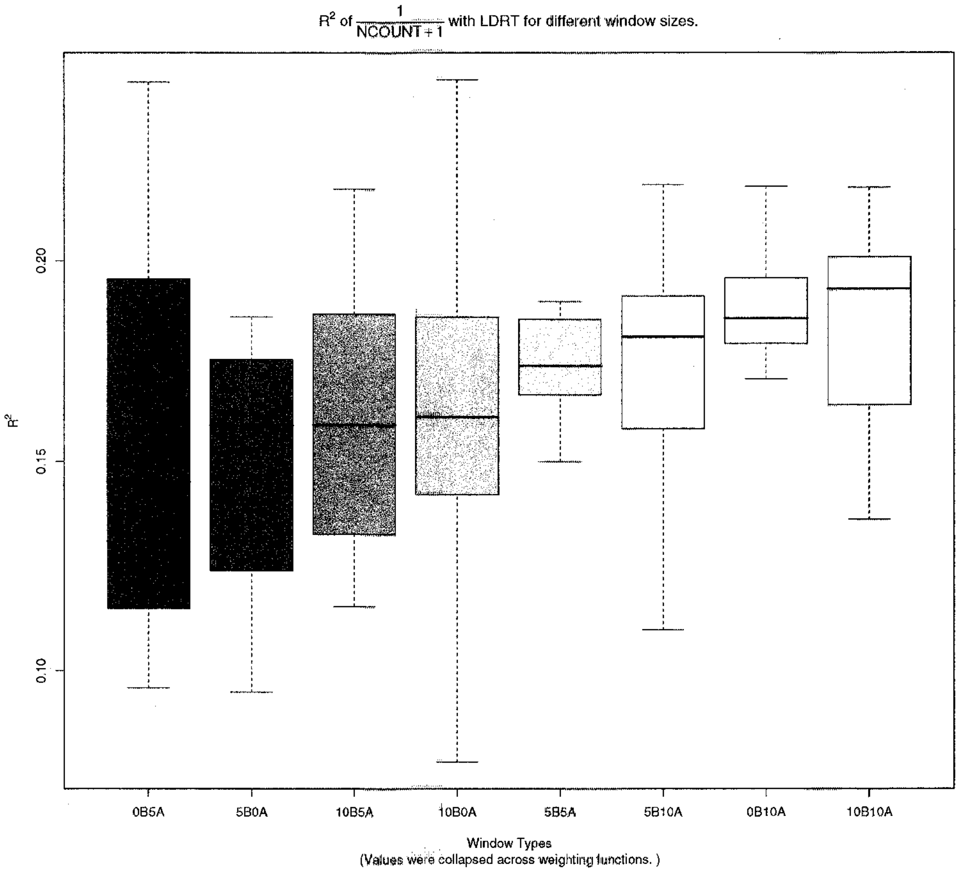


Figure 3.2: R^2 of NCOUNT-INV with LDRT, sorted by median R^2 for each window type.

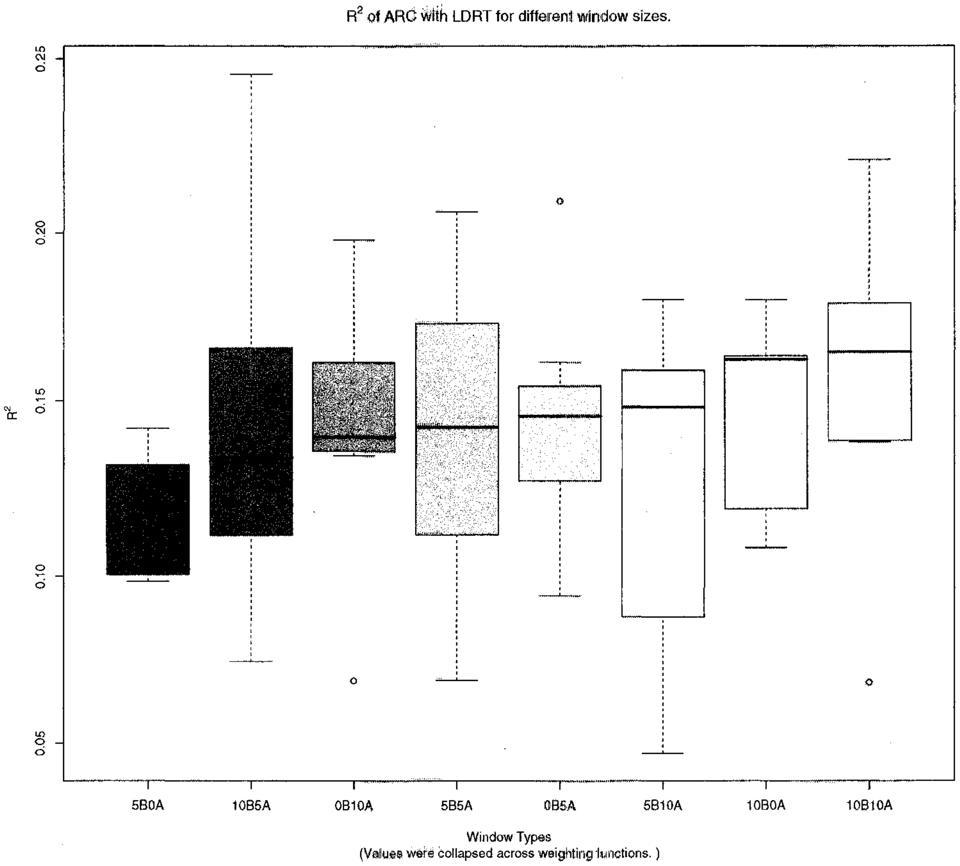


Figure 3.3: R^2 of ARC with LDRT, sorted by median R^2 for each window type.

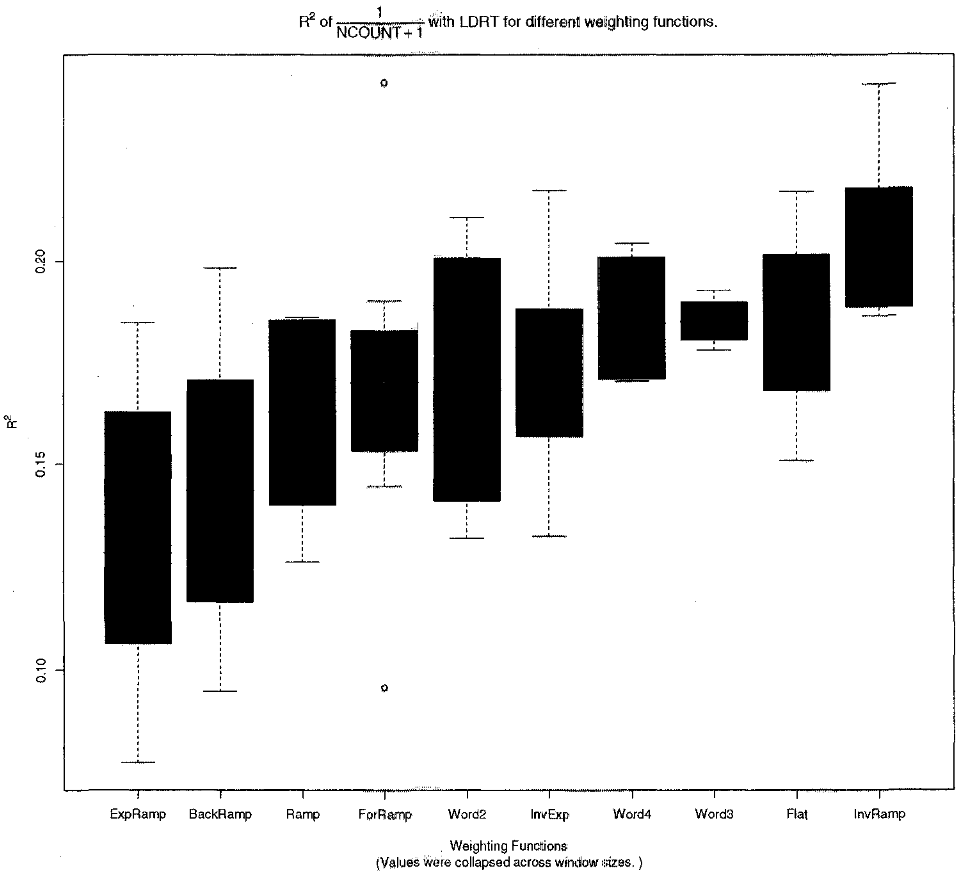


Figure 3.4: R^2 of NCOUNT-INV with LDRT, sorted by median R^2 for each weighting function.

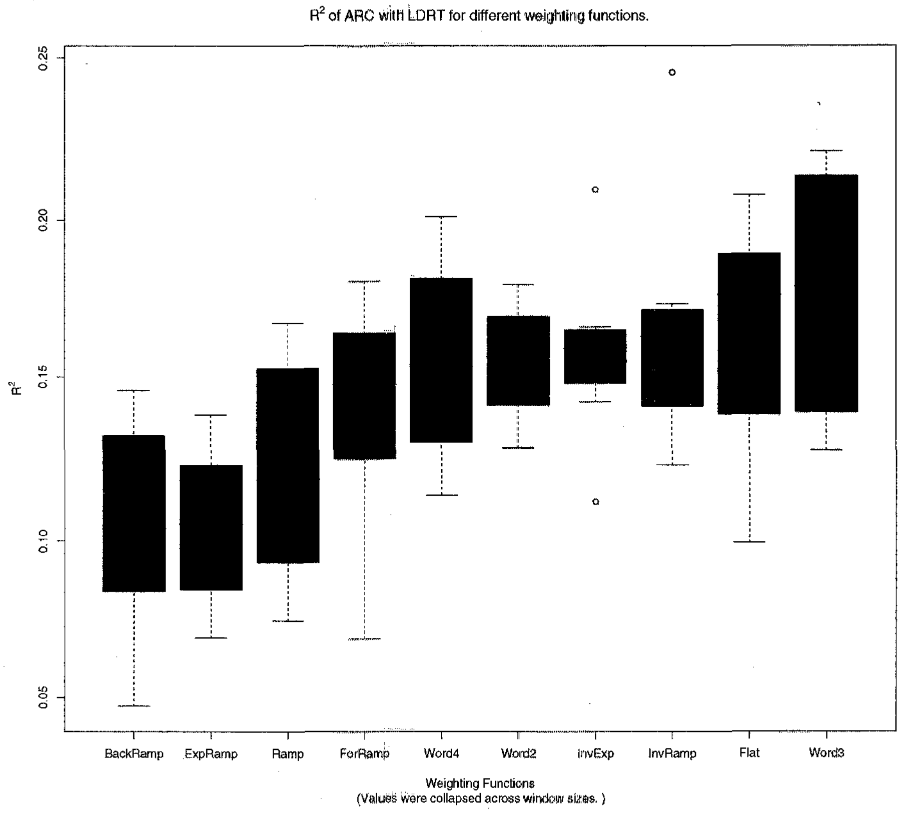


Figure 3.5: R^2 of ARC with LDRT, sorted by median R^2 for each weighting function.

To make the relative impact of window size and weighting scheme more easily understandable, all the data that is plotted in Figure 3.1 was sorted by highest median correlation, and then plotted in Figures 3.2, 3.3, 3.4 and 3.5. From these plots it appears that the best parameters judged by median correlation are 10B10A for window size, and Third Word or Inverse Ramp for weighting function.

The goal of this experiment was to explore the HAL parameter space and find the set of parameters that produced ARC and NCOUNT-INV measures that had the strongest correlation with experimental data from the English Lexicon Project (Balota et al., 2002). After much computation we found that our exploration of parameter space allowed us to achieve our first goal: to find out if these parameters make a difference or not. We found that there were large differences between the 73 parameter sets we tested. Do the parameters that we varied have a positive influence on the output of the model? By changing the window type and the window size we can change the R^2 between LDRT and ARC by up to 0.15. This implies that some parameter sets can explain up to 15% more of the variance than others, an extremely encouraging result. Certain parameters proved to be better at predicting LDRT than others, and these parameter sets deserve further study.

We found that when averaging across all the different weighting functions, the window size that consistently produced the greatest correlations with LDRT was the 10B10A, the window size used in HAL. However this window size did not produce the peak correlation, but rather 10B0A did. The combination of the 10B0A window size with the Fourth Word weighting scheme produced the highest correlation. The original HAL window size, 10B10A, was one of the best window sizes that we looked at.

As for the weighting scheme, the outcome was quite different. The original HAL weighting scheme, the Linear Ramp, fared poorly. It consistently had low correlations with LDRT. The Inverse Ramp weighting scheme performed consistently better. This may be due to the ability of the Inverse Ramp to give more weight to words that are further away in the window. These words can be more informative than the words that are directly adjacent which can often be closed class, function words. By de-emphasizing the minimal semantics of closed class words, the model may be improving the categorical relationships between words.

One constant across all of the results for these different parameter sets was the direction of the relationship between LDRT and ARC/NCOUNT-INV. The slopes produced by a linear regression of ARC or NCOUNT-INV on LDRT were uniformly greater than zero. In other words, the denser the neighborhood, the faster the reaction time. This type of facilitatory effect was reported by Buchanan et al. (2001). In experiment 4, Buchanan et al. (2001) used a factorial design (words with dense neighborhoods versus words with sparse neighborhoods) to investigate the relationship between neighborhood density and LDRT. They found that words with denser semantic neighborhoods had faster reaction times in a lexical decision task. The stimuli in this experiment were closely matched on orthographic neighborhood size and orthographic frequency, eliminating the influence of these lexical properties. The results presented here replicate the results of Buchanan et al. (2001) using a correlational design, and a larger set of stimuli (500 words versus 128 words) sampled from a larger pool of stimuli (32,000 mono- and multi-syllabic words versus 1,570 mono-syllabic words).

This experiment provided insight into the workings of HAL's parameters. To examine how this model could be applied to a task of greater psychological validity we moved beyond lexical decisions. We are bound by the shallow depth of semantic retrieval and processing that is inherent in LDRT data. In the next two experiments we will introduce a semantic decision task that we hope will help us gain greater insight into the representation and organization of lexical semantic memory.

3.3 Experiments 2A and 2B: A Semantic Decision Experiment

We devised two paired experiments to gather data relating to the semantic processing of words and cognitive load of a semantic decision task. Experiment 2A was a speeded forced choice semantic decision task that required participants to decide if words were related or unrelated. Experiment 2B was a judgment task where participants were asked to rate how related two words were. The design of this experiment was continuous and correlational: the stimuli were not separated into two categories for contrast. Rather, the stimuli were chosen to vary continuously over the range of HAL distances.

The aim of these experiment was to find out if there is a relationship between HAL measures of context distance and reaction times in a semantic task. The stimuli were chosen to represent a broad range of inter-word HAL distances. Our prediction was that the reaction times collected in experiment 2A would have a relationship with inter-word distance, and that the HAL measure produced by best parameter combinations found in experiment 1 would give better predictions than the default HAL parameters. Our intention was that the ratings collected in experiment 2B would allow us to compare the predictive power of subjective measures of relatedness with objective measures produced by HAL.

3.3.1 Subjects

64 undergraduate students enrolled in introductory psychology courses at the University of Alberta participated in this study for course credit (37 women, 27 men). Their mean age was 19.4 years old, and the standard deviation was 4.4 years.

3.3.2 Stimuli

300 pairs of words were chosen; 100 that were listed as associates in the Nelson Association Norms (Nelson, McEvoy, & Schreiber, 1998), 100 that were from the idiosyncratic (low frequency) responses list from the Nelson data, and 100 unrelated words. We built the stimuli sets with the goal of avoiding the dichotomy that is possible with semantically related word lists.

To make sure that our stimuli would cover a broad range of semantic relationships, we selected our stimuli very carefully. We built a very large set of pairs from which we chose smaller stimuli sets using a criterion of non-correlation. Two large sets of word pairs were all chosen from the full lists of word pairs from the University of Florida Nelson Norm databases. We started with the full list of 69,000 associated pairs (which we will call ASSOC) and the full list of 112,000 idiosyncratic responses (which we will call IDIO). The third large set was a list 200,000 word pairs that we generated ourselves by picking words randomly without replacement from a dictionary of English words with a frequency greater than 10 words per million (UNREL).

We measured the orthographic frequency (OF), orthographic neighborhood (ON) and word length (LEN) for all the words in these 287,000 pairs. We also calculated the inter-word distance in HAL space using the default HAL parameters. We then matched subsets of the three sets of word pairs so that for each ASSOC word pair, there would be an IDIO and an UNREL pair that were matched for OF, ON, LEN and HAL distance. The matching algorithm used was the following: all measures were converted into standard scores, and then the Euclidian distance between each pair and all the other pairs was calculated. The pairs with the smallest distance in z-score space were stored as a match and immediately removed from the input lists. This created three lists of approximately 1000 entries each, and from these lists, we picked the 300 pairs that satisfied the following criteria: pairs could not contain proper names, the UNREL pairs were not judged by either of two judges to be semantically related, and the HAL distance between all pairs was distributed evenly across the range of inter-word distance values.

To reduce the length of time that it took to participate in the experiment, we split the 300 pairs into two equal sets of 150 pairs (Parts X and Y) for use in Experiment 1 and 2 to counter balance the order of presentation. Equal numbers of participants did experiments 2A.X, 2B.X, 2A.Y and 2B.Y in both orders (Experiment 2A, then 2B, and vice versa).

The full stimuli set is available in Appendix A.

3.3.3 Method

In experiment 2A , stimuli were presented on an LCD display connected to a Macintosh computer (Mac OS X v. 10.3.9) using ACTUATE (Westbury, 2007). All words were displayed in lowercase letters in the Times Roman font. Participants were asked to make a judgment about two words that were to appear sequentially. The first word appeared in black at the top of a 500 pixel by 500 pixel white square for a duration of 2000 to 3500ms (this value varied randomly between all trials). Then, at the bottom of this square, a fixation point, the "+" symbol appeared for a duration of 500 to 1500ms (again, varying randomly), at which point the "+" symbol disappeared, and was replaced with the second word in the pair. This period of time, 2500 to 5000 ms, provided sufficient time for the participants to read the first word and access its meaning.

Once the second word appeared, participants were requested to make the following semantic decision (as explained in the instructions): "In your opinion, are these two words related?", as quickly and as accurately as possible. One of two keys on the computer keyboard, ("X" for No and "M" for Yes) were pressed, and the reaction time was measured.

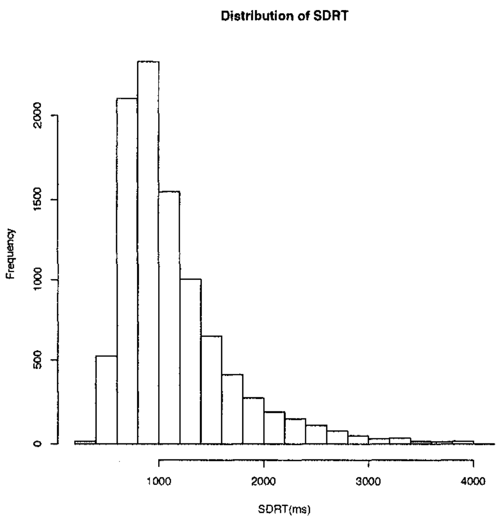


Figure 3.6: Distribution of 9524 RT oberservations.

For experiment 2B, participants were asked to do a slightly different task using the same apparatus and software as described for Experiment 2A. They were shown all the word pairs with both words appearing simultaneously, and were then asked to rate the relatedness of the words. The participants used a mouse to drag a sliding marker on a line on the screen. This line had the word "UNRELATED" over the left end of the line, and the words "HIGHLY RELATED" over the right side of the line. They were asked to take as much time as they needed to rate each pair of words. The software measured the position of the marker on the line and recorded 0 for UNRELATED and 100 for HIGHLY RELATED, as well as the time taken to do the rating.

3.3.4 Data Trimming and Analysis

We removed observations from experiment 2A that had an RT of less than 300ms or greater than 4500ms (two standard deviations from the mean). These outliers made up 1% of all observations. The distribution of reaction times, after removing outliers, is shown in Figure 3.6.

To do further analysis using measurements from our HAL model, we needed to calculate the mean RT for each item, but before we could calculate this statistic, we needed to remove data from trials where our participants made errors. We removed reaction times for ASSOC word pairs when the participant considered them unrelated. We also removed reaction times for UNREL items when the participant considered them to be related (see Table 3.2).

3.3.5 Results and Discussion

There was a strong effect of category on the reaction time that confirms that the stimuli were causing the desired pattern of cognitive load. As shown in Table 3.2, most of the ASSOC pairs were judged to be related in experiment 2A. The same held true for the IDIO pairs. For the UNREL pairs, they were mostly judged to be unrelated.

Category	**Semantic Decision**	**Number of obs. (% of total)**	**Mean SDRT (ms)**	**StdDev SDRT (ms)**
ASSOC	Related	2595 (81%)	1006.2	466.6
	Unrelated	575 (19%)	1210.1	544.5
IDIO	Related	2098 (66%)	1116.2	513.0
	Unrelated	1080 (34%)	1257.5	578.8
UNREL	Related	405 (13%)	1426.3	725.9
	Unrelated	2771 (87%)	1152.2	506.4

Table 3.2: Categorical distribution of the 300 Word Pairs with descriptive statistics

To assess the predictive power of the relatedness ratings in relation to the default HAL distances, we first analyzed the relationship between word ratings and reaction times.

For the relatedness ratings, we used the median rating for each word pair to analyze subjective relatedness. We did not use the mean so as to avoid the influence of extreme ratings. After a visual inspection of the scatterplot, we noticed a strong quadratic relationship (see Figure 3.7) between Mean SDRT and Median Relatedness Rating (MRR) for the 300 word pairs, characterized by the following regression equation, $R^2 = 0.40$, $F(2, 297) = 97.3, p < 0.001$:

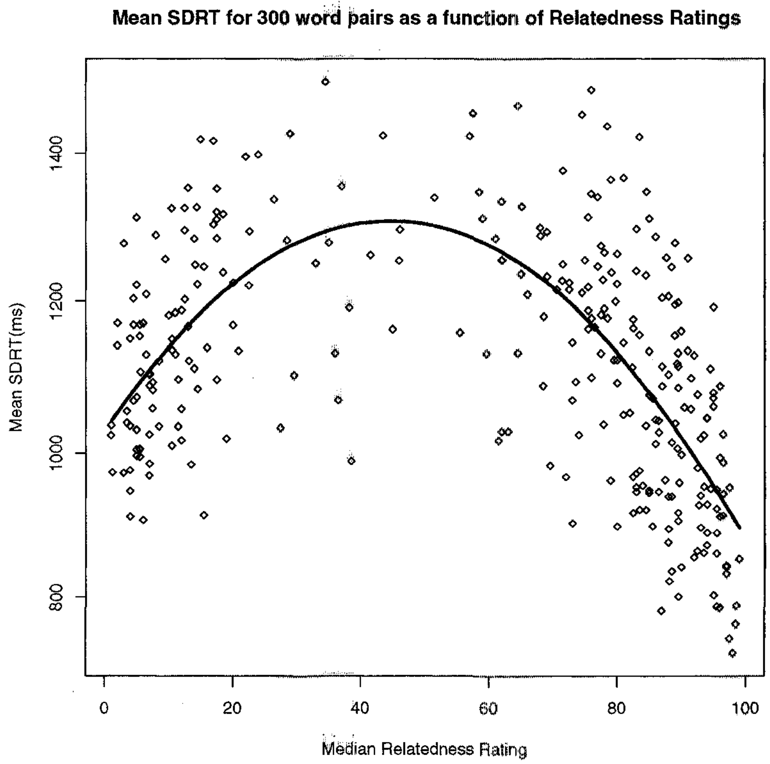


Figure 3.7: Mean SDRT for 300 word pairs as a function of relatedness ratings with quadratic least squares fit

$$\mu_{SDRT} = 1031.37 + 12.4\text{MRR} - 0.14\text{MRR}^2 \qquad (3.2)$$

This nonlinear, reverse U-shaped curve reconfirms our intuitive understanding of this semantic decision task. Words that are very unrelated or very related are quicker to process because they have meanings that are clearly convergent or divergent. Words that are rated as being weakly related have the slowest reaction times. This result highlights the difference between semantic priming in automatic retrieval/lexical decision versus semantic decision. In lexical decision semantic priming, unrelated words provide no facilitation of processing (Lucas, 2000). For this reason it is unwise to directly compare this semantic decision task with lexical decision semantic priming tasks.

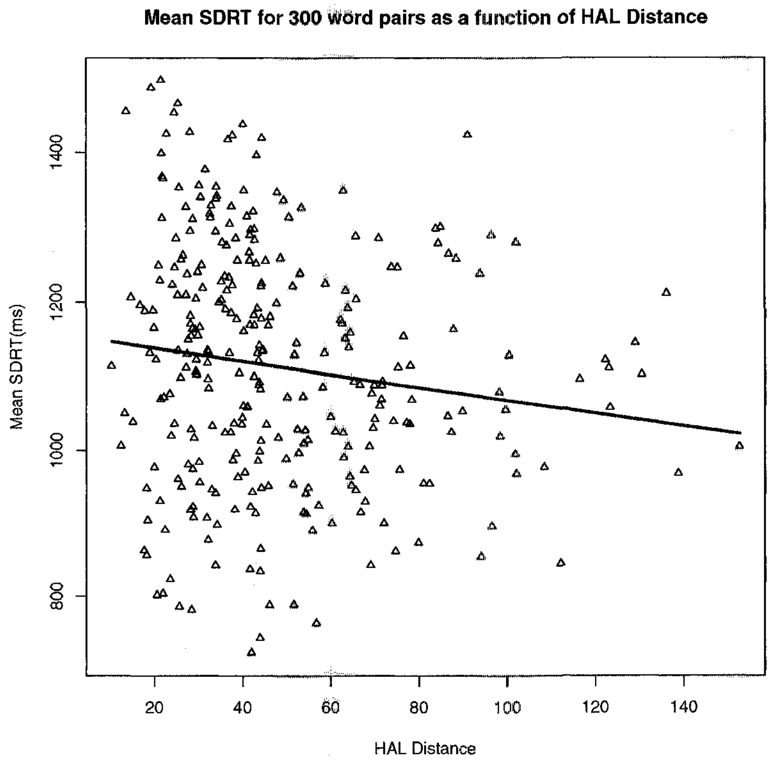


Figure 3.8: Mean SDRT for 300 word pairs as a function of HAL Distance

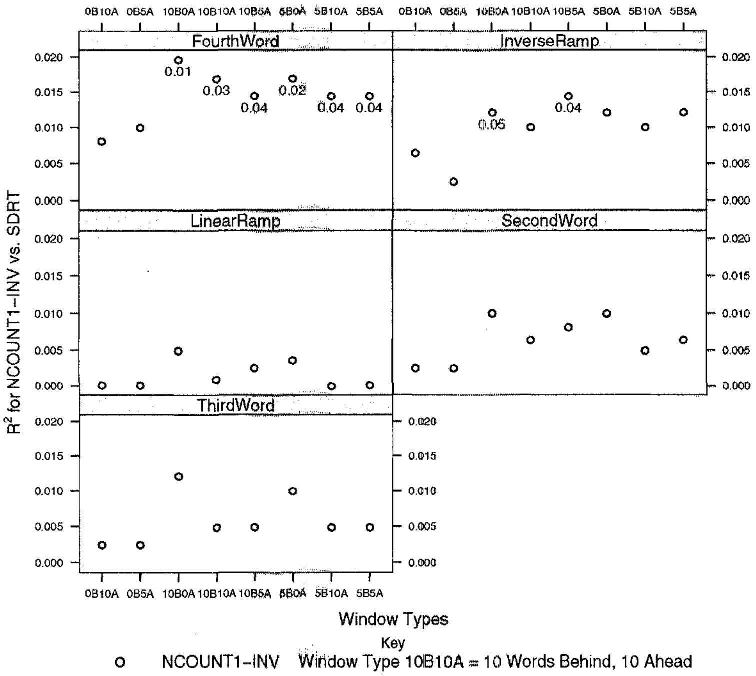


Figure 3.9: Overview of R^2 of NCOUNT-INV with SDRT for different weighting functions and window types. The p value for each reliable regression is shown below each point.

The increase in processing time from borderline words was seen by Vigliocco, Vinson, Damian, and Levelt (2002). They found that there there were graded semantic interference in a picture naming task depending on the semantic distance between the targets and distractors. They used a feature based semantic network with a small number of dimensions to calculate semantic distance.

The first HAL measure that we used in our analysis was the inter-word distance calculated with the default parameters from Lund and Burgess (1996). The linear regression for mean SDRT and this inter-word HAL distance was reliable but not strong, $R^2 = 0.016$, $F(1, 298) = 5.92, p < 0.02$ (see figure 3.8).

Using HiDEx, we calculated all the measures described in section 3.1.2 for all the word pairs in the stimuli set. We selected the most promising 40 parameter sets from the 73 parameter sets used in Experiment 1. These were all the parameter sets that included the following weighting schemes: Linear Ramp, Inverse Ramp, Second, Third and Fourth Word.

The only measure that had a significant correlation with SDRT was the semantic density for the first word, NCOUNT1-INV. In Figure 3.9 we see that the only two weighting functions to achieve statistically significant correlations were Fourth Word and Inverse Ramp. The most consistent window type was the 10B0A window. The best result was obtained from the combination of Fourth Word and 10B0A, $r = -0.14$, $R^2 = 0.02$, $F(1, 298) = 6.136$, $p = 0.01$. Since the slope of this relationship is negative, SDRT is predicted to decrease as the semantic density around the first word decreases (that is, as NCOUNT-INV increases). This result is congruent with the results from experiment 1 because the time required to access the meaning of the word pair and process the semantic information was less for words that had sparser neighborhoods.

In summary, we used a novel semantic decision task in Experiment 2 that allowed us to test predictions about the influence of HAL's parameters on reaction time. We presented two words, one after the other, and asked participants to decide if two words were related. We found that for small number of parameter sets, the density of the HAL neighborhood of the first word was a significant predictor of RT. The strength of the relationship was much less than that for the LDRT data in Experiment 1.

Why did our HAL measures explain so much less of the variability in this task than in the lexical decision task? One possibility is that there was mismatch between the questions we wanted to answer and the task we chose to use. The task that we used in this experiment was a very complex one. From the presentation of the first word until the presentation of the second word, the participants presumably accessed semantic information about the first word. When the second word is displayed, the reaction time measured will capture the time it takes to do at least two activities: retrieve the semantic information about the second item, and make a semantic decision. The SDRT we capture should be a function of the lexical retrieval of the second word, the complexity of the semantic memory traces for the two words, the type and number of relationships between the words, and the strategy/strategies that the participant used. We hypothesized that our HAL measures would provide an indirect measure of the complexity of the semantic traces and the relationship between the words (through their shared context). We have no data that will help us predict which strategies would be used to make these semantic decisions. This opens the door for variability that we will not be able to account for in our model.

Our concerns about the appropriateness of this forced choice task gave us the incentive to seek a better semantic decision task. This experiment and its results are described in the next section. We will reserve further discussion of the findings for the General Discussion.

3.4 Experiment 3: A Go/No-Go Semantic Decision Experiment

The task used in Experiment 2 was a forced-choice task ("Are these words related or unrelated?"). In Experiment 3 we attempted to replicate the semantic decision reaction time effect that we were interested in using a slightly different task. We chose a task that Siakaluk et al. (2003) found to be superior in eliciting semantic distance effects: the Go/No-Go semantic decision task. Siakaluk et al. (2003) used both a forced choice and a Go/No-Go task in a semantic decision experiment (Animacy: "Is this alive or not?"), and they noted that the time-constrained nature of the Go/No-Go task made it superior to other tasks for semantic decisions.

3.4.1 Subjects

35 undergraduate students enrolled in introductory psychology courses at the University of Alberta participated in this study for course credit (21 women, 14 men). Their mean age was 20.7 years old, and the standard deviation was 2.9 years. None of the subjects had participated in the previous experiments.

3.4.2 Stimuli

The stimuli used in this experiment were identical to those used in Experiments 2A and 2B.

3.4.3 Method

The laboratory equipment was identical to that used in Experiment 2. The only part of the procedure that was changed in Experiment 3 was the type of response that we requested of the participants after the second word appeared on the screen. The participants were instructed to press the space bar only if the words were related. If the words were unrelated, they were instructed to do nothing. If no input was detected after 3500ms, the next trial was initiated, and a No-Go result was recorded. 17 participants were show 150 pairs, and 18 participants were shown the remaining 150 pairs. Order of presentation was randomized for each subject.

3.4.4 Results and Discussion

The 35 participants performed a total of 5250 trials of which 57% were "Go" responses, and 43% were "No-Go" responses. We will only analyze the "Go" responses, as it is unclear how to interpret the lack of a response.

263 of the 300 word pairs were given at least one "Go" response. There is a strong possibility that some participants may have pressed the space bar hastily or unintentionally during the experiment. One way to detect unintended responses is

to look for items that produced very few "Go" responses. Any words with responses from less than 20% of the participants were removed (equivalent to 3 participants or less providing "Go" responses per item). After removing observations for these 56 items (3% of total number of observations), we analyzed the data for the remaining 207 word pairs.

Category	**Number of obs. (% of total)**	**Number of Word Pairs (% of total)**	**Mean SDRT (ms)**	**StdDev SDRT (ms)**
ASSOC	1565 (55%)	100 (49%)	1089.5	515.4
IDIO	1267 (43%)	94 (45%)	1249.6	573.8
UNREL	160 (2%)	13 (6%)	1812.8	783.3

Table 3.3: Categorical distribution of the 207 Word Pairs with descriptive statistics

As with the forced choice task, the ASSOC reaction times were the fastest, with the IDIO pairs having slower reaction times and greater variability. Due to the nature of the task, and non-responses, we were only able to collect reliable reaction times for a select few erroneously accepted UNREL words, and these reaction times were the longest and had the most variability.

To understand the relationship between the tasks in Experiments 2 and 3, we looked at the relationship for the mean SDRT for the 207 items that both experiments shared (see Figure 3.10). We found a very strong correlation between the logged reaction times of the two experiments $R^2 = 0.46$, $F(1, 205) = 172.4$, $p < 0.001$. We also found a strong quadratic relationship between the median relatedness ratings from Experiment 2 and the SDRT data from Experiment 3 $R^2 = 0.46$, $F(1, 204) = 88.19$, $p < 0.001$ (see Figure 3.11). There is a strong correspondence between the reaction times for items in Experiments 2 and 3. Across the experiments, there was no significant difference between the mean SDRT for the ASSOC stimuli ($t(198) = 1.195, p = 0.23$) or IDIO stimuli ($t(198) = 1.73, p = 0.09$). There was a significant difference of $386ms$ between the means for the UNREL stimuli in Experiment 2A and the means for the UNREL stimuli in Experiment 3

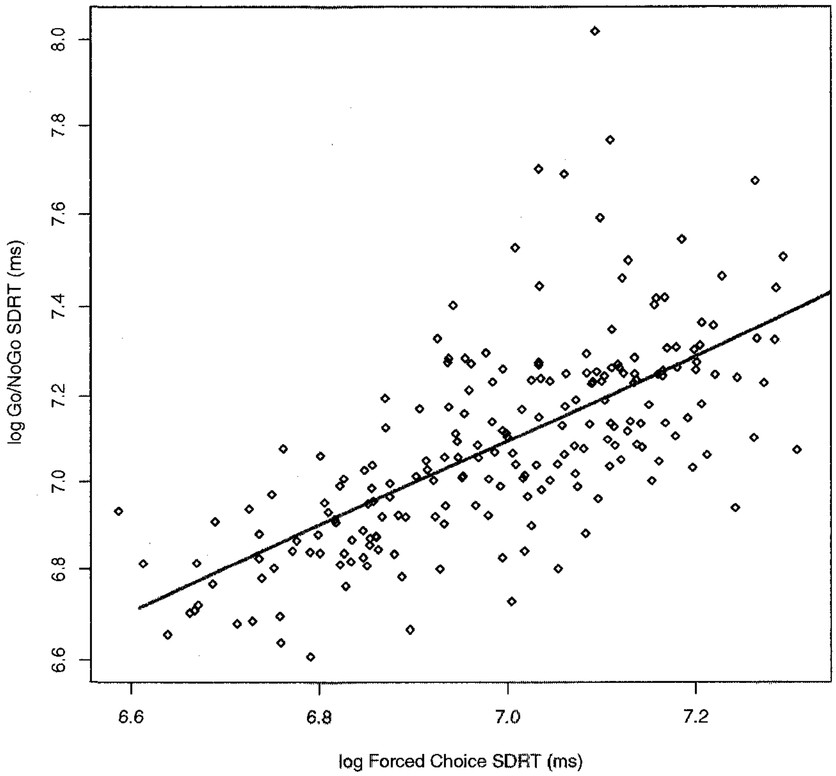


Figure 3.10: Log mean SDRT of the 207 Items in the Go/NoGo task as a function of Log mean SDRT in the Forced Choice task.

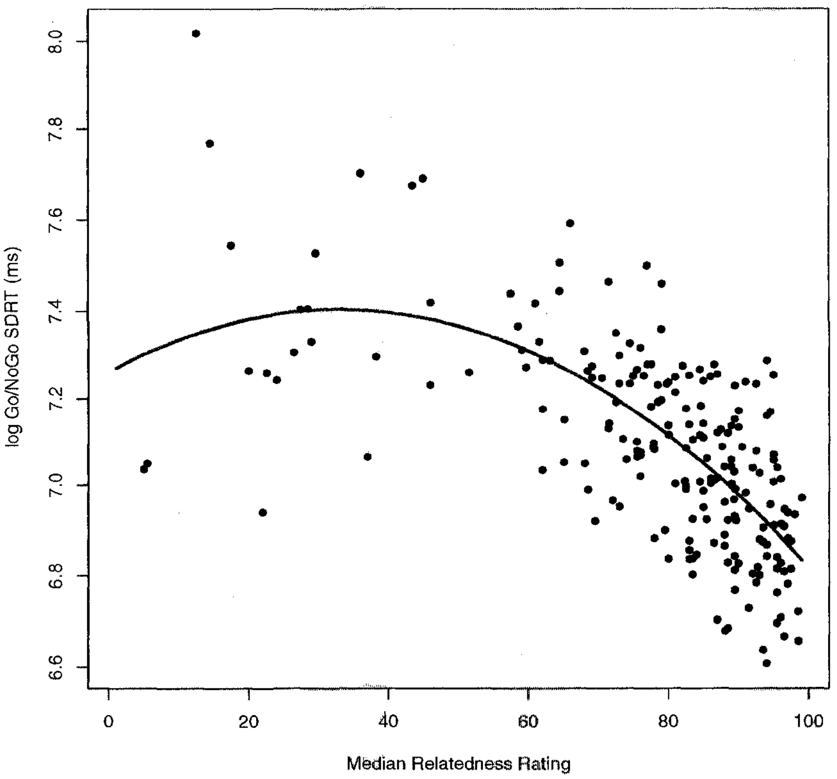


Figure 3.11: Log mean SDRT of the 207 Items in the Go/NoGo task as a function of median relatedness rating

($t(198) = 3.62, p < 0.001$), showing the predicted increased depth of processing for unrelated, difficult to process words.

The final step in the analysis is to study the relationship between SDRT in Experiment 3 and the measures calculated by HiDEx. We calculated the item regressions for mean SDRT and our seven measures for the same 40 parameter sets that were analyzed in Experiment 2. The results of this analysis are shown in 3.12. There was no significant correlation between mean item SDRT and the majority of the parameter sets. In particular, the original HAL parameters (10B10A, with the linear ramp weighting function) did not produce a significant correlation. As in Experiment 2, the only measure that had any significant correlation with a SDRT

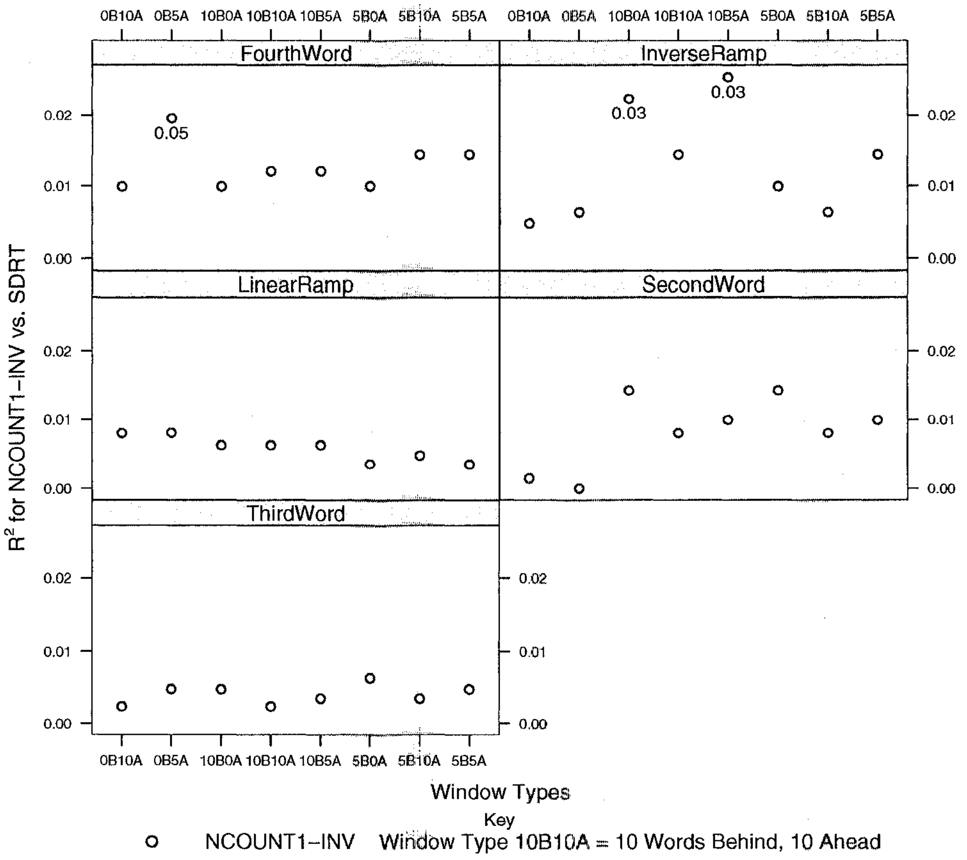


Figure 3.12: Overview of R^2 of NCOUNT-INV with SDRT in the Go/NoGo task for different weighting functions and window types. The p value for each reliable regression is shown below each point.

was NCOUNT-INV, the inverse of the number of neighbors plus one. The only two weighting functions that produced significant correlations with this measure were the Inverse Ramp and the Fourth Word functions. Both of these functions also performed well in our analysis of Experiment 2. The window types involved in the parameters sets that were 10B0A, 10B5A for the Inverse Ramp and 0B5A for Fourth Word.

The results obtained in Experiment 3 replicated the results from Experiment 2. There were three parameter sets that had significant linear relationships between neighborhood density and mean SDRT, and the amount of variance explained by these relationships was small. The implications of these results are discussed in the following section.

3.5 General Discussion

In Experiments 1, 2 and 3, we used LDRT and SDRT data to find optimal parameter sets for the HAL model. We found that certain weighting functions and window sizes fared much better than others at predicting reaction times. There was a clear consensus across all the experiments: the original HAL parameters do not create the best measures of neighborhood density for predicting lexical-semantic access time. There was an encouraging convergence in Experiments 2 and 3 that found that a small number of parameter sets produced the strongest correlations with a semantic decision task. The best weighting function for Experiments 2 and 3, the Inverse Ramp, was also the best for Experiment 1 (see Figure 3.4). The best window types were those that contained 10 word behind and 0 or 5 words ahead. These results suggest that there is a very important function served by the Inverse Ramp weighting function. What could be the reason for its superiority?

Unlike the Linear Ramp, greater significance is given to words that are located *further away* in the window from the word in question. This has the effect of reducing the impact of the words closest to the target word. What kinds of words are usually found in these positions? The intuitive answer is “function words” or “closed-class words”. Unlike nouns adjectives, adverbs and verbs, these are words cannot have any new members to their class (hence the “closed” class). They contain little semantic information about the words they appear next to, but do create semantic relationships between words in a sentence.

Are our intuitions correct about closed class words? In a cursory analysis of a 1 trillion word corpus of English culled from web pages (Brants & Franz, 2006) we found some interesting clues. Using a corpus-specific list of closed-class words of the 114 most frequent closed-class words in this corpus (see Appendix A.4 for this list), we counted the number of 2-grams in the corpus that contained at least one of these closed-class, ultra-high frequency words. We found that in the set of the 10,000,000 most frequent 2-grams, 50% of them contained at least one of these 114 words. This means that a very large proportion of the corpus is composed of 2-grams that contain closed-class words. Weighting schemes that reduce the weight given to closed-class word contexts may be better at capturing semantic context relationships because of the decrease in closed-class contextual information. This makes sense when we look at an example: the similarity between the contexts of "cat" and "dog" are more informed by "pet cat" and "pet dog" than by "the cat" and "the dog". By changing the weighting scheme, we changed the relative importance of closed-class word context, and made the model better.

Of particular note: in the semantic decision experiments, the neighborhood density of first word (and never the second word) in our word pairs produced the only significant relationships with SDRT. The denser the neighborhood of the first word, the longer the semantic decision took. This can only mean that the contextual richness of the first word is influencing the processing of the semantic decision, causing it to take longer. This relationship is in the opposite direction of the the relationship reported by Buchanan et al. (2001) and by Siakaluk et al. (2003). In these studies, a denser contextual neighborhood density was found to facilitate lexical access. A parallel phenomenon is seen in morphological family size (R. Baayen, Feldman, & Schreuder, 2006). There are also non-semantic neighborhood effects in lexical access, such as number of orthographic neighbors (Forster & Shen, 1996), that produce a similar effect of facilitation. In contrast, in our semantic decision experiments the opposite effect was seen: denser neighborhoods cause a slow down in reaction time. This type of competitive, or inhibitory, relationship has been found in some of the auditory lexical decision reaction time research as a function of the number of phonological neighbors (Luce & Pisoni, 1998). Our results indicate that a higher co-occurrence neighborhood density facilitates lexical access while simultaneously increasing the cognitive load of semantic decision processing.

The apparent contradiction between facilitatory and inhibitory effects of dense neighborhoods has recently been analyzed by Mirman and Magnuson (2007). They compared different models of semantic neighborhood density to find out if there were consistent facilitatory/inhibitory effects across different neighborhood density measures. They compared feature-based models, using data from Cree, McNorgan, and McRae (2006), association-space models, using data from Nelson et al., 1998, and co-occurrence models, using data from COALS (Rohde et al., 2007). They found that certain neighborhood measurements were correlated with facilitation while others were correlated with inhibition in both lexical decision and semantic decision tasks (living/non-living and abstract/concrete). In particular they found that a single measure of neighborhood density was unable to account for the pattern of results. Instead, they found that both the number of neighbors, and the distance of those neighbors was needed to understand the seemingly contradictory results. They reported that words with many near neighbors were categorized more slowly in a semantic decision task than words with few near neighbors. They also found that words with many distant neighbors were categorized in the same task more quickly than words with few distant neighbors. Mirman and Magnuson (2007) then go on to model this phenomenon with a feature vector based attractor model (a type of neural network model; see Cree et al., 2006 for the model's architecture). In light of this work by Mirman and Magnuson (2007), we propose an alternative interpretation of our results from Experiments 2 and 3: since our NCOUNT-INV measure is built using a threshold, and it only counts the nearest neighbors, it is also capturing data about how many nearby neighbors a word has. Independently, we have found an identical inhibitory effect for neighborhood density to the one that was found by

Mirman and Magnuson (2007), despite the fact that we used a relatedness judgment task while they used other semantic categorization tasks.

How does this result relate to previous research into lexical semantic processing? Our semantic decision task is unlike most semantic psycholinguistic tasks. An extensive amount of research has been done on semantic priming (for a review, see: Moss & Tyler, 1995). Unlike most semantic priming experiments, this semantic task we developed was not a lexical decision task. There is no implicit, subliminal semantic activation. The facilitation or inhibition in our experiments were the result of a combination of the semantic relationship between the words in the pair and the participant's strategies. This difference in methodology makes comparisons of effect size between our experiments and lexical decision semantic priming experiments difficult. What about semantic categorization/semantic decision tasks? The difference between traditional semantic decision tasks and our tasks is that in most semantic decision tasks, a category, such as "concrete words", or an exemplar, such as "an animal", are used throughout the experiment. The task for the participant is usually a category membership decision that stays constant throughout the experiment. In our task, the category or exemplar is different in each and every trial. This makes it difficult to compare our results with with those from traditional semantic decision experiments. For example, it may explain why our reaction times are much slower than those in experiments which used for semantic decisions for concreteness (Binder, Westbury, McKiernan, Possing, & Medler, 2005) and animacy (Siakaluk et al., 2003).

There is at least one study that used a task very similar to ours (a forced choice relatedness task) to study semantic processing. Pexman, Hino, and Lupker (2004) used a relatedness task to investigate ambiguity in semantic processing. They found that for "no" trials (trials where the two words were not related) there was no ambiguity effect, and on related trials, there was an ambiguity disadvantage. They also proposed that this disadvantage was due to the semantic decision task itself, and not the process of retrieving the semantic representations for the words. We did not collect ambiguity ratings for the words in our stimuli set. This makes it difficult to compare our experiments with those in Pexman et al. (2004). The relationship between ambiguity and co-occurrence neighborhood density merits further study.

There are two potential concerns about our experimental paradigm, controlling imageability and amount of variability explained. We attempted to control for many psychologically relevant lexical variables in Experiments 2 and 3, but there is (at least) one variable that we were unable to control for that has been shown to influence reaction times. It is imageability or concreteness (see reviews by Paivio, 1991; Schwanenflugel, 1991). We were unable to add this variable into our stimuli selection process because there were insufficient numbers of words with published imageability ratings to make our stimuli set. There is evidence that imageability may influence how words are processed (Binder et al., 2005), and future work should include more control of stimulus imageability.

A final concern is the small amount of variability explained by our HAL measure, NCOUNT-INV for the first word, in Experiments 2 and 3. Even with the best parameter settings, only 2% of the variability is explained by our context density measure. The meaning of this number needs to be clarified. It is the amount of variability attributable to NCOUNT-INV after controlling for all the other lexical variables described in Section 3.3.2. No other model of lexical semantic memory or processing has ever been used to model this type of task. This amount of correlation is small but reliable. Perhaps more variability can be explained in the future with more investigation into the structure of lexical memory and other models of lexical memory.

Chapter 4

Conclusions

High-dimensional models of word meaning are very powerful psychological models of memory, and their vectorial representation of information has a strong theoretical foundation in the work of Churchland (1989). By linking neurobiology with vector computation, Churchland's framework opens the door to using vectors to represent our memory of word meaning. The conclusions that follow are based on the idea that psycholinguistic phenomena can be explained by models that perform computations on data from high-dimensional spaces. HAL is one such model, and we found that there are reliable relationships between the output of a HAL-like model and experimental data from human subjects.

We have presented three experiments that explore the effect of varying two HAL parameters on modeling semantic processing tasks. Experiment 1 compared the lexical decsion reaction time predictions of 73 different HAL parameter sets. Experiments 2 and 3 used, respectively, a forced choice task and a Go/NoGo task, to see if a task with an increased semantic load would show a predictive pattern for the HAL model. We found that, for certain optimized parameter sets, ARC and NCOUNT-INV were able to account for a large amount of variability in lexical decision reaction times. We then tested the power of these near-optimal parameter sets to predict semantic decision reaction time in a novel task. The amount of variability explained by the optimal parameter sets in the semantic decision model was small in comparison to the lexical decision model, but converged on the same parameter settings that were found in the LDRT experiments. We have shown that changing the weighting function and window size parameters of the HAL model can have a powerful impact on the ability of HAL to predict LDRT and SDRT. Additionally, the best set of parameters found were *not* those used in the original HAL model by Lund and Burgess (1996).

Finally, we found that the best set of parameters for predicting reaction times were convergent for the SDRT and LDRT data. This finding opens the door to more research using HAL as a model for predicting behavioral data. Linguistic tasks that have a large semantic component could be modeled with a HAL-like representation of semantic information. If these models had their best fit using the same parameters discussed here, it would point to a general applicability of these parameter settings.

What are we doing when we change these HAL parameters? In a very broad sense, we are tuning the input to our vector representations. As Churchland (1989) noted, the input given to a high-dimensional vectorial representation will largely determine its output. Just as a better-shaped ear will filter out noise and improve auditory representations, better lexical context representations for HAL will produce better semantic representations of the words in the model. We will consider the impact of our parameter tuning on the input as well as the vectors produced from that input to make our conclusions.

We have looked at how the local co-occurrence frequencies are weighted before being input into the model. The optimal weighting schemes, Inverse Ramp and Fourth Word, reduce the influence of the words directly preceding or following a word in its context. We analyzed a very large corpus of English, and found that the vast majority of adjacent words are closed-class words. We speculate that these closed-class words can act like "noise" in our model, whereas the contextual information in the open-class words nearby are the "signal". In practical terms, the co-occurence values for closed-class words will be smaller relative to open class when using the Inverse Ramp or Fourth Word weighting schemes. The weighting scheme that allowed us to best predict behavior are the ones that filter out information about co-occurrence with closed-class words.

If the weighting scheme parameter has a potential psycholinguistic link, what about the window size parameter? Why is the optimal window setting that we found, 10B0A, better than the others? The relative importance of the backward window over the forward window might be due to the way that working memory stores recently perceived language. Only the most recently heard words are kept in the phonological loop (Baddeley, 2003) in much the same way that only the most recently seen words are kept in the 10B0A window. Furthermore, specific language impairment (SLI) has been linked to impairments of working memory, and suspected to lead to problems in acquiring the meaning of words (Baddeley, 2003). If the concept "working memory span" can be considered analogous to the idea of "window size", then perhaps the optimum size of a person's working memory span for learning the meaning of words can be modeled using HAL. Removing the influence of preceding words removes half the information from a the original HAL global co-occurrence matrix, shrinking the actual dimensionality, and therefore the size of high-dimensional space. The benefits of finding solutions in a space with less dimensions may be coming into play here.

The potential connections between HAL's weighting scheme, HAL's window size and psychological theories demonstrate the continued relevance of HAL-like memory models to the theoretical debates about the structure of lexical-semantic memory. Furthermore, these explanations fit nicely into Churchland's (1989) framework of vectorial computation. The weighting scheme may improve the quality of information in the vector representation, and the window size may remove an unnecessary portion of the vector representation from the model.

4.1 Other models

Are distributed representations the only choice for representing lexical semantics, or are there other viable models? HAL (Lund & Burgess, 1996), COALS (Rohde et al., 2007) and BEAGLE (Jones & Mewhort, 2007) are all highly distributed representations. Some neural network models, such as the SRN model proposed by Elman (2004) and the feature-based model by Cree et al. (2006) are also distributed word representations. In contrast, there are also localist semantic representations, often called "lexico-semantic networks". These networks were originally described by Collins and Loftus (1975) as part of their spreading activation model, and inspire much of the current research that involves semantic networks built from WordNet (Fellbaum, 1998), which is a handmade data-set of semantic relationships. WordNet-based distance metrics and neighborhood density measures have been used to predict LDRT, but have produced lower correlations than distributed representations (Rohde et al., 2007). For this reason we did not include an analysis of the predictive power of a competing localist representation in this work. There is still much work to be done before we can intelligently compare the relative merits of local and distributed representations.

4.2 Future Work

There is much more work to be done based on the results we obtained. The experimental paradigm of the Go/NoGo task could be improved by changing the default, NoGo response to be for RELATED words. This change would have the effect of making participants look for the absence of relationships, a profoundly different task, and perhaps one that would be well modeled by HAL. Another methodological improvement could be to use only one exemplar for a set of contiguous trials instead of changing the exemplar on every trial. This would allow the participant to calibrate their criteria for one exemplar and avoid any task confusion.

There may be ways to take the ideas introduced in this research program in new directions. For example, the relatedness task in Experiments 2 and 3, seen from a slightly broader viewpoint, is similar to a conceptual combination task. We could ask future participants: "Do these two concepts combine well?" Future work could use our high-dimensional models as semantic representation that could be included in models of conceptual combination, such as CARIN (Competition Among Relations in Nominals) theory (Gagne & Shoben, 1997).

Beyond representations lies the question of lexical processing. Our research dealt specifically with two processes that occur concurrently with lexical access and retrieval: word recognition (Experiment 1) and relatedness recognition (Experiments 2 and 3). Future research could involve looking at models that deal with semantic information in word processing, and how our results are informed by these models. Many models of word recognition contain a "semantic" layer, component or module. The majority of these word recognition models do not impose a requirement on which type of semantic representation is to be used with the model. We were unable to find any reports of any attempts to compare the impact of local versus distributed models of representation on word recognition models, but many connectionist models of word recognition assume a distributed model of representation. The most relevant models are the Interactive Activation Model (Rumelhart, 1981), the MultiStage Cascade model (Borowsky & Besner, 1993), the Independent Activation Model (Dixon & Twilley, 1999a , Dixon & Twilley, 1999b, Twilley & Dixon, 2000) and the Single Mechanism Model (Plaut & Booth, 2000). All of these models have the potential to incorporate HAL-based semantic vectors into their semantic module, and until such work is undertaken there will no way to see how our representation will perform in these models.

Even more fundamental is the question of language acquisition. HAL is a model of representation of semantic knowledge about words. It is by nature a statistical model which keeps track of first and second order co-occurrence probabilities. An open question remains: does the statistical nature of the representation imply a statistical learning paradigm? What are people doing when they learn new words? They may be looking for statistical patterns in their input. Some initial work on using statistical patterns to model language acquisition through neural networks (which are mathematically equivalent to HAL) has been done by Howell, Jankowicz, and Becker (2005). There is some reason to speculate that language acquisition does depend on statistical learning, as seen the the work of Saffran, Aslin, and Newport (1996) on phonology acquisition by 8 month old infants. There is much work ahead in order to understand the acquisition of lexical semantic knowledge.

Making HAL and HAL-like models more psychologically plausible is a noble goal, but a difficult one. HAL does not use any information about word morphology, phrase structure, or any other non-lexical linguistic information. HAL requires large amounts of electronic text to function properly. What is truly fascinating is that despite its inherent simplicity, HAL can model human performance on complicated semantic tasks fairly well.

References

Baayen, R., Feldman, L., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, *55*, 290-313.

Baayen, R. H. (2001). *Word frequency distributions.* Boston, MA, USA: Kluwer Academic.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX Lexical Database [CD-ROM].* Philadelphia, PA USA: Linguistic Data Consortium.

Baddeley, A. (2003, 0). Working memory and language: an overview. *Journal of Communication Disorders,*, *36*(3), 189-208.

Balota, D., Black, S. R., & Cheney, M. (1992, May). Automatic and attentional priming in young and older adults: Reevaluation of the two-process model. *Journal of Experimental Psychology: Human Perception and Performance*, *18*, 485-502.

Balota, D., Cortese, M., Hutchison, K., Neely, J., Nelson, D., Simpson, G., et al. (2002). *The English Lexicon Project: A web-based repository of descriptive and behavioral measures for 40,481 English words and nonwords.* Web resource. Retrieved October 5th, 2005, from: `http://elexicon.wustl.edu/`, Washington University, St. Louis, MO.

Balota, D., & Spieler, D. (1999). Word frequency, repetition, and lexicality effects in word recognition tasks: Beyond measures of central tendency. *Journal of Experimental Psychology: General*, *128*, 32–55.

Barsalou, L. W., Simmons, W. K., Barbey, A. K., & Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Sciences*, *7*, 84-91. (Cited By (since 1996): 71)

Binder, J. R., Westbury, C. F., McKiernan, K. A., Possing, E. T., & Medler, D. A. (2005). Distinct brain systems for processing concrete and abstract concepts. *Journal of Cognitive Neuroscience*, *17*, 905-917.

Borowsky, R., & Besner, D. (1993). Visual word recognition: A multistage activation model. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*, *19*, 813-840.

Brants, T., & Franz, A. (2006). *Web 1T 5-gram Version 1.* Philadelphia, PA USA: Linguistic Data Consortium.

Buchanan, L., Burgess, C., & Lund, K. (1996). Overcrowding in semantic neighborhoods: Modeling deep dyslexia. *Brain and Cognition*, *32*, 111–114.

Buchanan, L., Westbury, C., & Burgess, C. (2001). Characterizing semantic space:

Neighborhood effects in word recognition. *Psychonomic Bulletin & Review, 8*, 531-544.

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

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*, 188-198.

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

Burgess, C., Livesay, K., & Lund, K. (1998). Explorations in context space: words, sentences, discourse. *Discourse Processes, 25*, 211–257.

Burgess, C., & Lund, K. (1997). Modelling parsing constraints with high-dimensional context space. *Language & Cognitive Processes, 12*, 177–210.

Burgess, C., & Lund, K. (2000). The dynamics of meaning in memory. In E. Dietrich & A. B. Markman (Eds.), *Cognitive dynamics: Conceptual and representational change in humans and machines* (pp. 117–156). Mahwah, NJ, US: Lawrence Erlbaum Associates.

Churchland, P. (1989). *A neurocomputational perspective: the nature of mind and the structure of science.* Cambridge, MA: MIT Press.

Churchland, P., & Sejnowski, T. (1992). *The computational brain.* Cambridge, MA: MIT Press.

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

Cree, G. S., McNorgan, C., & McRae, K. (2006). Distinctive features hold a privileged status in the computation of word meaning: Implications for theories of semantic memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 643-658.

Dixon, P., & Twilley, L. C. (1999a). Context and homograph meaning resolution. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie experimentale, 53*, 335-346.

Dixon, P., & Twilley, L. C. (1999b). An integrated model of meaning and sense activation and disambiguation. *Brain and Language, 68*, 165-171.

Elman, J. (1993). Learning and development in neural networks: The importance of starting small. *Cognition, 48*, 71–99.

Elman, J. (2004). An alternative view of the mental lexicon. *Trends in Cognitive Sciences, 8*, 301–306.

Fellbaum, C. (Ed.). (1998). *Wordnet: An electronic lexical database.* Cambridge, MA: Bradford Books.

Forster, K. I., & Shen, D. (1996). No enemies in the neighborhood: Absence of inhibitory neighborhood effects in lexical decision and semantic categorization. *Journal of Experimental Psychology: Learning, Memory and Cognition, 22*, 696–713.

Fristrup, J. A. (1994). *USENET: Netnews for everyone.* Englewood Cliffs, NJ, USA: Prentice Hall.

Gagne, C. L., & Shoben, E. J. (1997). Influence of thematic relations on the comprehension of modifier-noun combinations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 71-87.

Harris, Z. (1968). *Mathematical structures of language.* New York: Wiley.

Howell, S. R., Jankowicz, D., & Becker, S. (2005). A model of grounded language acquisition: Sensorimotor features improve lexical and grammatical learning. *Journal of Memory and Language*, *53*, 258-276.

Jackendoff, R. (1983). *Semantics and cognition.* Cambridge, MA.: MIT Press.

Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution.* New York: Oxford University Press.

Jackendoff, R. (2007, 5/18). A parallel architecture perspective on language processing. *Brain Research*, *1146*, 2-22.

Jones, M. N., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, *55*, 534-552.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*, 1-37.

Lucas, M. (2000). Semantic priming without association: A meta-analytic review. *Psychonomic Bulletin and Review*, *7*, 618–630.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear & Hearing*, *19*, 1–36.

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

Mirman, D., & Magnuson, J. S. (2007). Attractor dynamics and semantic neighborhood density: Processing is slowed by near neighbors attractor dynamics and semantic neighborhood density: Processing is slowed by near neighbors and speeded by distant neighbors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *in press*.

Moss, H., & Tyler, L. (1995). Investigating semantic memory impairments: the contribution of semantic priming. *Memory*, *3*, 359–395.

Murdock, B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, *89*, 609–626.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). *The university of south florida word association, rhyme, and word fragment norms.* (http://www.usf.edu/FreeAssociation/)

Osgood, C. E., Suci, G., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana, IL, USA: University of Illinois Press.

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal oF Psychology*, *45*, 255–287.

Pexman, P. M., Hino, Y., & Lupker, S. J. (2004). Semantic ambiguity and the process of generating meaning from print. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*, *30*, 1252-1270.

Plaut, D. C., & Booth, J. R. (2000). Individual and developmental differences in semantic priming: Empirical and computational support for a single-mechanism account of lexical processing. *Psychological review, 107*, 786-823.

Rohde, D. L. T., Gonnerman, L. M., & Plaut, D. C. (2007). *An improved method for deriving word meaning from lexical co-occurrence.* Unpublished manuscript. Retrieved April 20th, 2007, from: `http://tedlab.mit.edu/~dr/`, Massachusetts Institute of Technology, Cambridge, MA.

Roy, D., Patel, R., DeCamp, P., Kubat, R., Fleischman, M., Roy, B., et al. (2006). The human speechome project. In *In proceedings of the 28th annual cognitive science conference.* Mahwah, New Jersey, USA: Lawrence Erlbaum.

Rumelhart, . J. L. M., D. E. (1981). Interactive processes in reading. In C. Perfetti & A. Lesgold (Eds.), (chap. Interactive processing through spreading activation.). Hillsdale NJ, USA: Erlbaum.

Russell, B. (1910). *The study of mathematics: Philosophical essays.* London: Longmans, Green.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*, 1926 - 1928.

Schwanenflugel, P. (1991). Why are abstract concepts hard to understand? In P. Schwanenflugel (Ed.), *The psychology of word meanings.* (p. 223-250). Hillsdale, NJ, US: Lawrence Erlbaum Associates.

Shaoul, C., & Westbury, C. (2006). Word frequency effects in high-dimensional co-occurrence models: A new approach. *Behavior Research Methods, 38*, 190-195.

Shaoul, C., & Westbury, C. (2007). *A usenet corpus (2005-2007)* (Tech. Rep.). Edmonton, AB: University of Alberta. Downloaded from http://www.psych.ualberta.ca/ westburylab/downloads/usenetcorpus.download.html.

Siakaluk, P. D., Buchanan, L., & Westbury, C. (2003). The effect of semantic distance in yes/no and go/no-go semantic categorization tasks. *Memory & Cognition, 31*, 100-113.

Song, D., Bruza, P., & Cole, R. (2004). *Concept learning and information inferencing on a high-dimensional semantic space.* ACM SIGIR 2004 Workshop on Mathematical/Formal Methods in Information Retrieval (MF/IR'2004), 30 July 2004, Sheffield UK.

Song, D., Bruza, P., Huang, Z., & Lau, R. Y. (2003). Classifying document titles based on information inference. In J. G. Carbonell & J. Siekmann (Eds.), *Foundations of intelligent systems* (p. 297-306). Berlin/Heidelberg, Germany: Springer.

Song, D., & Bruza, P. D. (2001). *Discovering information flow using a high dimensional conceptual space.* The 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (New Orleans, LO).

Twilley, L. C., & Dixon, P. (2000). Meaning resolution processes for words: A parallel independent model. *Psychonomic Bulletin and Review, 7*, 49-82.

Vigliocco, G., Vinson, D., Damian, M., & Levelt, W. (2002). Semantic distance effects on object and action naming. *Cognition, 85*, B61–B69.

Westbury, C. (2007). *Actuate: The alberta something or other.* (Under Review. Downloaded from `http://www.psych.ualberta.ca/~westburylab/`)

Wittgenstein, L. (1958). *Philosophical investigations.* Englewood Cliffs, NJ, USA: Prentice Hall. (Third edition. Translated by G.E.M. Anscombe.)

Yates, M., Locker, L., & Simpson, G. B. (2003). Semantic and phonological influences on the processing of words and pseudohomophones. *Memory and Cognition, 31*, 856-866.

Zipf, G. K. (1935). *The psycho-biology of language.* New York: Houghton-Mifflin.

Zipf, G. K. (1949). *Human behaviour and the principle of least-effort.* New York: Addison-Wesley.

Appendix A

Stimuli used in Experiment 2

A.1 Associated Words sorted by Inter-word HAL distance

WORD PAIR	HAL DISTANCE	OF (Word1)	OF (Word2)	LETTERS (Word1)	LETTERS (Word2)	ON (Word1)	ON (Word2)
ESSAY - ENGLISH	41.97	6.58	93.71	5	7	1	0
FACTORY - LABOR	42.15	21.42	38.80	7	5	1	1
ARREST - SUSPECT	42.55	20.87	42.56	6	7	1	0
DIFFER - SIMILAR	43.79	9.30	110.97	6	7	2	1
AVERAGE - REGULAR	43.89	66.71	52.13	7	7	1	0
BIOLOGY - CELL	43.99	11.80	43.69	7	4	0	16
CELL - BIOLOGY	43.99	43.69	11.80	4	7	16	0
LEGS - STRETCH	44.11	24.48	10.13	4	7	15	0
BOTHER - TROUBLE	44.20	29.37	64.90	6	7	4	0
EXPERTS - PANEL	44.21	32.77	21.73	7	5	3	2
DECENCY - RESPECT	44.40	16.87	65.89	7	7	1	0
WORRY - PANIC	45.87	37.31	8.72	5	5	4	1
KNIGHT - HORSE	46.09	11.39	24.55	6	5	0	10
CORRUPT - LAWYER	48.11	21.84	27.26	7	6	0	1
KINGDOM - QUEEN	49.96	27.20	22.20	7	5	0	3
FAKE - PRETEND	50.20	22.34	18.81	4	7	16	1
LIQUID - DRINK	51.62	10.81	29.06	6	5	0	4
FENCE - CHAIN	54.20	11.57	32.74	5	5	2	1
ENGINE - LOUD	54.84	45.61	17.12	6	4	0	6
DOCTOR - OPERATE	55.88	45.30	23.32	6	7	0	1
CLOUD - WEATHER	56.76	27.77	24.91	5	7	2	3
SPORTS - STADIUM	57.23	24.41	18.39	6	7	7	0
COUNSEL - LAWYER	58.17	14.17	27.26	7	6	0	1
PIPE - VALVE	59.95	15.14	7.33	4	5	9	6
COWBOY - RANGE	60.20	5.76	70.58	6	5	1	3
DEPART - AIRPORT	61.01	47.06	20.05	6	7	1	0
DEPTH - WIDTH	63.90	15.75	87.42	5	5	2	0
BEACH - RELAX	64.69	29.81	5.61	5	5	7	2
CORRUPT - DESTROY	65.74	21.84	40.24	7	7	0	0
ARTIST - TALENT	66.76	14.10	19.96	6	6	0	0
RECIPE - MIXTURE	67.82	10.07	8.12	6	7	1	1
MERCURY - PLANETS	69.05	17.50	11.22	7	7	1	1
OPENING - VALVE	69.80	49.03	7.33	7	5	1	6
FANTASY - PRETEND	71.08	20.84	18.81	7	7	2	1
FORTUNE - FAME	71.54	23.52	9.70	7	4	0	15
DELIVER - TRUCK	72.02	32.08	22.17	7	5	1	4
LICENSE - PERMIT	74.70	36.26	14.27	7	6	1	1
COURAGE - COWARD	77.16	12.84	16.95	7	6	0	2
BONDAGE - CHAIN	79.88	5.05	32.74	7	5	1	1
LAUNCH - MISSILE	81.01	20.95	14.85	6	7	2	3
DOORWAY - PORTAL	82.36	14.02	14.40	7	6	0	3
ECONOMY - DEFICIT	88.44	47.30	10.45	7	7	0	0
WINDOW - SHIELD	91.07	62.29	8.65	6	6	1	0
ENGLISH - POETRY	94.08	93.71	11.75	7	6	0	0
CHICKEN - RECIPE	96.47	17.67	10.07	7	6	1	1
ANCIENT - TEMPLE	101.83	29.78	19.87	7	6	1	0
MAILBOX - EMPTY	102.09	7.01	65.53	7	5	0	1
WORSHIP - TEMPLE	112.10	27.05	19.87	7	6	1	0
HELPFUL - USELESS	136.19	30.30	18.69	7	7	0	0
SESSION - THERAPY	138.77	21.90	12.60	7	7	1	0
Means:	**66.56**	**26.24**	**28.60**	**6.24**	**6**	**2.22**	**2.06**

A.2 Idiosyncratic Words sorted by Inter-word HAL distance

WORD PAIR	HAL DISTANCE	OF (Word1)	OF (Word2)	LETTERS (Word1)	LETTERS (Word2)	ON (Word1)	ON (Word2)
DRAMA - SERIOUS	29.48	7.44	96.40	5	7	1	0
SHARP - HARSH	29.76	54.36	10.53	5	5	4	1
NOTION - PURPOSE	32.91	18.00	67.18	6	7	4	0
ROUTINE - CYCLE	34.71	21.68	39.60	7	5	1	1
FOOLISH - ACTIONS	36.50	11.77	63.83	7	7	1	0
MESS - KITCHEN	38.56	21.12	13.40	4	7	15	0
MONDAY - BUSY	42.46	43.81	24.58	6	4	0	6
PLASTIC - TREND	42.58	21.00	13.83	7	5	3	2
APPLES - TREES	42.86	5.73	36.47	6	5	0	10
PRIEST - COLLEGE	43.54	20.68	112.91	6	7	2	1
UNITE - APART	44.43	5.49	32.41	5	5	2	1
GRAVITY - ROCK	46.24	18.23	44.05	7	4	0	16
LUCK - RAINBOW	47.83	46.22	6.92	4	7	16	0
PARK - RESERVE	48.67	39.98	15.60	4	7	16	1
SELECT - CAREFUL	49.48	48.43	25.04	6	7	0	1
WILD - BEAST	51.50	28.48	12.66	4	5	9	6
PRAISE - DESTROY	51.52	16.85	40.24	6	7	1	0
SUCCEED - LUCKY	52.19	13.58	28.90	7	5	0	3
CITIZEN - ARREST	52.45	30.55	20.87	7	6	0	1
IMAGINE - SUPPOSE	53.20	52.52	42.64	7	7	1	0
PANEL - BUTTONS	58.59	21.73	26.69	5	7	2	3
FRIGID - WORRY	58.90	5.04	37.31	6	5	0	4
EXPRESS - HURRY	62.19	70.76	6.63	7	5	1	6
MORALS - IMMORAL	62.93	5.06	8.26	6	7	7	0
PILOT - ERROR	62.95	18.09	90.28	5	5	2	0
ADVISE - CONSOLE	63.44	12.27	9.26	6	7	1	1
COLLECT - MESS	64.02	16.63	21.12	7	4	0	15
SUCCEED - WEALTH	64.35	13.58	26.69	7	6	0	1
BREAD - FLOUR	65.67	14.23	5.57	5	5	7	2
CEMENT - FLOOR	67.61	5.66	57.32	6	5	1	3
ETHICS - VIRTUE	69.22	17.97	9.78	6	6	0	0
CURIOUS - MYSTERY	69.95	18.85	14.33	7	7	2	1
PASSAGE - MYSTERY	70.93	15.04	14.33	7	7	1	1
REALIZE - BEAUTY	71.48	46.44	15.89	7	6	1	1
POVERTY - STUDENT	71.65	19.08	38.64	7	7	0	0
REGION - PORTION	73.86	48.12	19.97	6	7	1	0
SEVERE - WEATHER	75.25	18.96	24.91	6	7	2	3
SERVANT - CASTLE	78.03	7.24	19.66	7	6	0	2
CAPTURE - STEAL	83.82	17.23	17.07	7	5	1	4
FAILURE - DISMISS	84.92	51.42	6.76	7	7	0	0
DEFEND - KNIGHT	86.49	35.53	11.39	6	6	1	0
CAPTURE - RESCUE	86.77	17.23	14.63	7	6	1	1
KINGDOM - MICKEY	87.74	27.20	6.94	7	6	0	3
ANCIENT - MAGIC	89.82	29.78	29.47	7	5	1	1
FREEDOM - ESCAPE	98.18	99.08	19.53	7	6	0	0
COLLECT - ITEMS	100.43	16.63	55.93	7	5	0	1
CAPTURE - VICTIM	102.03	17.23	29.58	7	6	1	0
KINGDOM - BRITAIN	123.04	27.20	31.07	7	7	0	0
DECENCY - ETHICS	123.29	16.87	17.97	7	6	1	0
SURGERY - MIRACLE	128.94	17.14	8.74	7	7	1	0
Means:	**65.55**	**25.46**	**28.88**	**6.24**	**6**	**2.22**	**2.06**

A.3 Unrelated Words sorted by Inter-word HAL distance

WORD PAIR	HAL DISTANCE	OF (Word1)	OF (Word2)	LETTERS (Word1)	LETTERS (Word2)	ON (Word1)	ON (Word2)
FRUITS - TOWARDS	28.68	5.77	112.61	6	7	2	1
SELDOM - VOTES	33.06	6.12	24.64	6	5	0	10
BLUNT - REACT	36.97	44.69	11.75	5	5	4	1
FOUGHT - OBVIOUS	37.78	18.63	63.98	6	7	4	0
HELMET - QUOTE	37.98	6.27	68.23	6	5	1	3
RICE - ELEMENT	38.51	33.01	20.37	4	7	16	1
BELOVED - QUERY	39.73	8.73	38.49	7	5	1	1
SLOWLY - TUBE	40.49	43.56	14.47	6	4	0	6
PAYROLL - DIETS	41.07	25.31	21.29	7	5	0	3
ESSENCE - PARK	41.57	11.51	39.98	7	4	0	16
INVOLVE - SCHEME	42.71	19.59	28.70	7	6	0	1
WIND - SIGNALS	43.39	27.73	15.36	4	7	15	0
HINT - SANDY	43.74	13.00	8.04	4	5	9	6
CIRCUIT - ASPECTS	44.02	22.77	38.59	7	7	0	0
BROKEN - REMARKS	44.59	39.11	17.89	6	7	1	0
WITNESS - BROAD	50.64	23.22	22.78	7	5	3	2
MISS - PATRIOT	51.81	39.46	13.47	4	7	16	0
KNEES - ENDED	52.78	8.90	30.91	5	5	2	1
TREATY - ANIMALS	53.46	11.75	43.82	6	7	1	0
PROPHET - MONTHLY	53.83	24.85	68.38	7	7	1	0
DIVORCE - GOTTEN	53.91	10.60	25.83	7	6	0	1
VOTED - KNEES	54.37	24.82	8.90	5	5	7	2
LINKED - POPCORN	54.89	18.01	15.41	6	7	7	0
JUICE - PERFORM	62.49	9.40	94.95	5	7	1	0
LEMON - ELECTED	62.78	21.93	27.41	5	7	2	3
AWHILE - SUITE	63.21	7.92	33.44	6	5	0	4
FINEST - SYMBOL	63.96	10.78	17.62	6	6	0	0
SMOOTH - CLARIFY	64.37	11.84	7.56	6	7	1	1
APPEARS - FAILURE	65.22	61.30	51.42	7	7	1	0
COMPARE - COAST	65.88	24.68	25.98	7	5	1	4
ARRIVES - PENALTY	66.52	20.49	16.98	7	7	2	1
FEMALE - ENTERED	68.80	45.40	23.39	6	7	0	1
CRUELLY - EXPLORE	69.59	22.46	9.80	7	7	1	1
PACIFIC - DOZENS	74.17	14.58	17.84	7	6	0	3
JOINED - CRYSTAL	75.24	23.91	17.75	6	7	2	3
SHUTTLE - FAVOR	75.60	6.10	33.87	7	5	1	1
PACIFIC - PACE	76.42	14.58	11.95	7	4	0	15
IMAGINE - BENCH	77.85	52.52	7.24	7	5	1	6
KNOCK - THIRD	78.32	12.46	90.50	5	5	2	0
ENFORCE - STREAM	84.38	9.51	14.04	7	6	0	2
CLASSIC - EXPAND	87.31	28.89	16.19	7	6	1	1
FRIENDS - GUARDS	93.91	96.63	12.71	7	6	0	0
PACKAGE - MIRRORS	96.41	45.88	8.27	7	7	0	0
MILITIA - UPPER	98.37	11.51	67.46	7	5	0	1
BEHAVE - POETRY	99.60	53.28	11.75	6	6	1	0
ELEMENT - SCREAM	108.35	20.37	8.93	7	6	1	1
ORDERED - ENIGMA	116.38	32.55	19.53	7	6	1	0
SOLDIER - FORUMS	122.25	20.54	19.75	7	6	1	0
WEEKEND - TACTICS	130.50	30.34	18.66	7	7	0	0
CHICKEN - DEPOSIT	152.59	17.67	14.28	7	7	1	0
Means:	**66.41**	**24.30**	**29.06**	**6.24**	**6**	**2.22**	**2.06**

A.4 114 Most Frequent Closed-Class Words from the Web1T Corpus

A	ABOUT	ALL	ALSO
AN	AND	ANY	ARE
AS	AT	BE	BEEN
BUT	BY	CAN	DO
FIRST	FOR	FROM	GET
HAD	HAS	HAVE	HE
HER	HERE	HIS	HOW
I	IF	IN	INTO
IS	IT	ITS	JUST
LIKE	MAKE	MAY	ME
MORE	MOST	MY	NO
NOT	NOW	OF	ON
ONE	ONLY	OR	OTHER
OUR	OUT	OVER	SHOULD
SO	SOME	SUCH	THAN
THAT	THE	THEIR	THEM
THERE	THESE	THEY	THIS
THROUGH	TO	UP	US
WAS	WE	WERE	WHAT
WHEN	WHICH	WHO	WILL
WITH	WOULD	YOU	YOUR
$< /S >$	$< S >$	@	=
>	?	’	’S
(	)	*	+
,	-	–	.
...	/	:	;
[	\	]	|
!	“	#	$
%	&		

All of these words were among the 200 most frequently used words in the Web1T corpus. $< S >$ and $< /S >$ denote the beginning and end of a sentence.