What is Semantic Distance? A Review and Proposed Method for Modeling Conceptual Transitions in Natural Language

Jamie Reilly[1,2], Bonnie M. Zuckerman[1,2], Ann Marie Finley[1,2],
Celia P. Litovsky[1,2], & Yoed N. Kenett[3]

[1]Eleanor M. Saffran Center for Cognitive Neuroscience
[2]Department of Communication Sciences and Disorders
Temple University, Philadelphia, Pennsylvania USA

[3]Faculty of Industrial Engineering & Management
Technion – Israel Institute of Technology, Haifa, Israel

Address Correspondence to:
Jamie Reilly, Ph.D.
Temple University
1701 N. Cecil B Moore Avenue
Philadelphia, PA 19122
reillyj@temple.edu
Tel: (215) 204-3996

Abstract

Cognitive science has seen a rapid evolution of the tools available for studying conceptual knowledge. Historically, much of our understanding of semantic memory has been informed through studies using language. Natural language processing (NLP) has offered groundbreaking techniques for elucidating relationships between concepts and language at unprecedented scales. One of the most popular applications has involved mathematically representing concepts using high dimensional feature spaces. Here we describe the nature of such semantic spaces and review ways in which human- and machine-generated semantic distance metrics differ in capturing taxonomic (e.g., *dog-wolf*) versus thematic (e.g., *dog-leash*) semantic relationships. We propose a novel method and open-source algorithm for deriving semantic distances between adjacent content words in connected language samples. This R package transforms a user-specified language transcript into a vector of pairwise semantic distances spanning all adjacent bigrams (e.g., The cat drank the milk → cat-drink, drink-milk, etc.). These distances constitute a continuous time series reflecting word-by-word level changes in meaning across a language sample of any length. We derive semantic distance norms and apply the proposed technique to a classic work of short fiction, *To Build a Fire* (Jack London, 1908). We discuss extensions of this time series approach, including the potential for forecasting, causal modeling, topic cohesion, and as an implicit measure of semantic impairment in spoken and/or written narratives.

1. Introduction

Semantic memory operates both at the microscale level in representing the meanings of individual concepts and at the macroscale level when constructing meaning between concepts (Hills & Kenett, 2022; Jones et al., 2015; Kumar, 2021). Much of our understanding of conceptual knowledge has been informed through language-based paradigms involving production and/or comprehension of single words or meticulously controlled arrays of words. This pattern is evident across a wide range of experimental tasks such as blocked cyclic naming, semantic decision, lexical decision, priming, and picture-word interference (Binney et al., 2018; Capitani et al., 2003; Cutler, 1981; Farah & McClelland, 1991; Funnell et al., 2006; Grossman et al., 2004; Hillis & Caramazza, 1991; Hodges et al., 1996; Kousta et al., 2011; Lupker, 1979; Pexman et al., 2016; Warrington, 1975; Woollams et al., 2008). Advantages gained in experimental control can, however, come at a cost to ecological validity. People do not communicate using single words. Language is an emergent system whose elements often combine in non-linear and unpredictable ways to convey meaning at different levels of discourse (Price et al., 2015; Westerlund & Pylkkänen, 2014).[1]

---

[1] A principled argument exists for delineating words from concepts (Malt, 2020; Malt et al., 2011; Malt & Majid, 2013). Although many past works have used the terms 'word' and 'concept' synonymously, it is important to note this distinction. The semantic models we review here are language-based but generally agnostic to this conceptual-lexical distinction.

Techniques leveraged from natural language processing (NLP) and machine learning have recently facilitated more widespread use of connected discourse (e.g., storybooks, podcasts, corpora) in studies of semantic processing (Baldassano et al., 2017; Deniz et al., 2019; Günther et al., 2019; Hartung et al., 2020; Huth et al., 2016; Jain & Huth, 2018; Kumar, 2021; Mandera et al., 2015, 2017; Nastase et al., 2020; Popham et al., 2021a). A key advantage of such approaches is their capacity to extract distributional statistics from vast numbers of tokens appearing in real world corpora (Baldassano et al., 2017; de Heer et al., 2017; Hartung et al., 2020; Huth et al., 2012; Jain & Huth, 2018; Naselaris et al., 2011; Popham et al., 2021b; Simony et al., 2016).

The growing popularity of NLP-based approaches may be attributable to their capacity to quantitatively represent semantic distances between words (Kumar et al., 2021). Such metrics have proven useful for modeling phenomena such as semantic priming (Kenett et al., 2017; Kumar et al., 2020), as well as complex dynamics such as associative thinking (Beaty et al., 2021; Gray et al., 2019) and creative thinking (Beaty & Johnson, 2021; Kenett, 2018, 2019; Olson et al., 2021).

1.1 What is semantic distance?

Semantic distance is a quantitative measure of relatedness between two or more concepts. Consider, for example, the relationship between dogs and wolves, two natural kinds within the genus *Canis* that share a high feature overlap (e.g., body morphology,

kinship behaviors). If the constraints of such a taxonomy are known, one might readily derive a measure of similarity between dogs and wolves. Such Linnean taxonomies based largely on physical similarities are one schema among an infinite field of both well-defined (e.g., domesticated/wild) and ad hoc distinctions (e.g., cuddliness) (Chrysikou, 2006; Barsalou, 1983, 1991). Semantic distance is accordingly a flexible rather than fixed construct. Two concepts might be closely related on one dimension but distantly related across countless others.

Recent work has situated word meaning within high dimensional feature spaces using both human (e.g., crowd sourcing) and machine learning (i.e., word embedding) algorithms (Binder et al., 2016; Crutch et al., 2013; Fernandino et al., 2016; Landauer & Dumais, 1997; Prabhakaran et al., 2014; Reilly et al., 2016; Troche et al., 2017). Figure 1 illustrates a simple two-dimensional semantic space bounded by two factors, affective valence and physiological arousal. It is possible to situate any two words (e.g., hatred vs. love) within such a space. In this particular example, we derived Euclidean distance between *hatred* and *love* using the Warriner and colleagues psycholinguistic norms for arousal and valence (2013).[2]

---

[2] This dataset contains crowdsourced Likert-scale ratings for arousal and valence of 13,915 English lemmas, including *hatred* (valence = 2.38, arousal = 5.22) and *love* (valence = 8.00, arousal = 5.36).
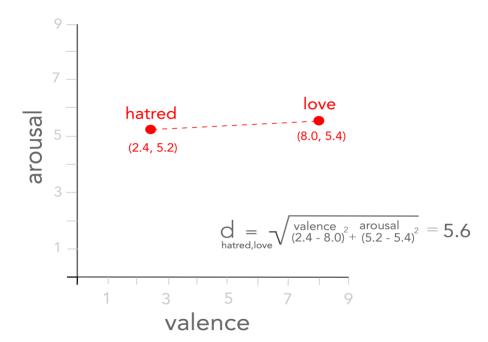
Figure 1. Semantic Distance between Love and Hatred

$$d_{hatred,love} = \sqrt{\underset{valence}{(2.4 - 8.0)}^2 + \underset{arousal}{(5.2 - 5.4)}^2} = 5.6$$

1.2 Taxonomic vs. Thematic Semantic Systems

Cognitive science has converged upon several broad conceptual distinctions in constraining the possible range of semantic spaces. The optimal semantic space has both neurobiological plausibility as well as explanatory power for a wide range of behavioral, neuropsychological, and neuroimaging phenomena (Beauchamp et al., 2002; Binder et al., 2016; Crutch et al., 2013; Devlin et al., 1998; Kalénine et al., 2009; Martin, 2015; Mirman et al., 2017; Reilly et al., 2021; Sacchett & Humphreys, 1992). One underlying assumption of all semantic space models is that word meaning is decompositional. That

is, concepts are not unitary constructs but are instead the aggregate of a wide array of co-activated features (e.g., color, pleasantness) (Allport, 1985; Patterson, 2007; Rosch, 1973).

Semantic feature processing represents one of the most widely investigated domains in classical theories of concept representation (Dilkina & Lambon Ralph, 2012; Gonnerman et al., 1997; Jackendoff, 1987; McNorgan et al., 2011; McRae et al., 1999; Mirman et al., 2017). Feature-listing approaches (e.g., has a___, is a___ , used for__) have been especially informative for specifying hierarchical semantic relationships that mark dogs and wolves as close taxonomical relatives (Cree & McRae, 2003; Garrard et al., 2001; McRae et al., 1999; Rogers et al., 2004). Two recent semantic space models exemplify such feature-based or taxonomic perspectives. Binder and colleagues proposed an experiential model where target words are characterized along 65 discrete sensorimotor, affective, and interoceptive dimensions (2016). Crutch and colleagues proposed an analogous approach (i.e., abstract conceptual feature space) characterizing words along twelve dimensions (Crutch et al., 2013; Reilly et al., 2016; Troche et al., 2017). Both models rely on crowd-sourced data, wherein participants provided Likert-scale ratings of their own subjectively perceived salience for hundreds of words on each dimension (e.g., color, sound).

1.2 Context and co-occurrence as unique domains in semantic memory

"A man is known by the company he keeps" ~ Aesop (620 – 564 BCE)

Aesop's proverb highlights a core tenet of human social cognition. We often make wide-ranging assumptions about unfamiliar people by evaluating the characteristics of their friends. In language networks, 'friends' are words that regularly co-occur with a given target word. Thus, people can analogously learn much about words by looking to their neighbors (Landauer & Dumais, 1997). For example, dogs, collars, bones, frisbees, and leashes are all semantically bound through a common event structure. We learn through repeated exposure that the conditional probability of encountering a collar in the context of a dog is high. In turn, dogs and collars are linked by the crucial distinction of "what goes together" in semantic memory (Hoffman et al., 2011, 2018; Sadeghi et al., 2014). Such contextually bound associations are maintained by the thematic semantic system (Geller et al., 2019; Kalénine et al., 2009; Maki & Buchanan, 2008; Mirman et al., 2017; Schwartz et al., 2011).

Co-occurrence and embedding are key concepts underlying thematic knowledge. Forks and paprika, for example, are both embedded within kitchens despite sharing no sensorimotor features. Reliable co-occurrence among 'kitchen things' scaffolds an associative semantic network based around a particular schema. Such co-occurrence statistics about the world arrive through numerous modalities (sensorimotor, affective, interoceptive), including language.

Language-based co-occurrence reflects the embedding of two or more units (words, n-grams, sentences) within a given temporal window. Word embedding models such as LSA (Landauer & Dumais, 1997), Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), ELMo (Peters et al., 2017), BERT (Devlin et al., 2018), and GPT-2 (Radford et al., 2018) hold a profound assumption about language co-occurrence:

1.3 Words that appear together often are semantically related.

One of the most radical implications of word embedding models is their concrete mechanism for implicit conceptual learning. Knowledge of a new or unfamiliar word can be at least partially induced by indexing that word's neighbors. Consider a person who no knowledge of the meaning of *funeral.* She might gradually learn about funerals through exposure to various linguistic contexts in which *funeral* is embedded (e.g., "Funerals are so <u>sad</u>.", "The <u>bereaved widow</u> was <u>crying</u> at the funeral.", "The funeral was the <u>worst</u> day of my life."). The neighbors of *funeral* (i.e., sad, bereaved, crying, worst, widow) constitute an associative semantic network. This network allows listeners to bootstrap the meanings of unknown words.

Our ability to draw effective inferences from associative semantic networks relies on breadth and stability of the network. We gradually associate funerals with sadness through extensive and consistent exposure to language. Computers are also remarkably adept at detecting co-occurrence patterns gleaned from large corpora (i.e., collections of

structured text such as books, newscasts, transcribed podcasts, or Twitter feeds). Most word embedding approaches first tokenize a corpus into a vector of unique words. Next, a large matrix is constructed representing the count of each unique word within a specified moving window size. Since most words *never* co-occur within the moving window specified, these matrices tend to have high sparsity, necessitating dimension reduction techniques that ultimately vectorize the meaning of each word into a relatively small number of hyperparameters.

Unlike experiential semantic models vectorized by features (e.g., color, disgust), hyperparameters derived from word embeddings are abstract mathematical constructs. The representation of any token (e.g., word, n-gram, phrase) in word embedding space reflects its unique values distributed across hyperparameters. Semantic distance between two words is typically reported as the cosine of the angle between two vectors ranging from -1 to 1 (Landauer & Dumais, 1997; Mikolov et al., 2013; Pennington et al., 2014).[3]

1.4 Study Aims

---

[3] The angle between two identical vectors (e.g., any word plotted against itself) is 0° , and the cosine of 0 is equal to 1. Therefore, a cosine distance of 1 reflects the highest possible similarity between any two words. -1 reflects the highest dissimilarity.

Word embedding and experiential semantic models are yielding insights into the neural correlates of semantic processing, creativity, and the integrity of semantic networks in neuropsychiatric conditions (Anderson et al., 2019; Beaty et al., 2021; Beaty & Johnson, 2021; Fernandino et al., 2016, 2022; Hills et al., 2015; Hills & Kenett, 2021). Little is known about how semantic distance manifests in connected language. To follow, we propose a novel method of deriving word-by-word semantic distances in continuous narratives. We introduce an open-source algorithm that transforms language transcripts into vectors of temporally ordered words, followed by computation of all running pairwise distances (bigram distances). We also report norms for taxonomic and thematic semantic distances, derived by querying millions of words from contemporary English texts. Finally, we demonstrate how inter-word semantic distance can provide a sensitive metric of cohesion among the flow of ideas in a renowned work of short fiction, *To Build a Fire* (Jack London, 1908).

2 Methods

2.1 Overview

To follow, we outline: a) a text-cleaning algorithm; b) a novel taxonomic semantic feature space; c) a thematic semantic space derived from a word embedding model trained on a large corpus of written English; d) procedures for generating semantic distance norms across many adjacent word pairs in running language; e) illustrative

applications of these procedures to a work of fiction. All scripts and data are available for download and use at https://osf.io/ryhfj/.

2.2 Text Cleaning Algorithm

Our aim here was to prepare any user-specified text for semantic distance calculations. We developed a cleaning algorithm that first applied numerous global omissions and substitutions (e.g., words, numbers, non-alphabetic characters, stop-words) followed by lemmatization of the remaining content words. This script implements many individual regular expressions and wrapper functions to read, strip, and format text (see Table 1). Readers are invited to inspect all individual commands by visiting the R-package vignette for 'semdistflow' (Zuckerman, Reilly, Litovsky, & Finley, 2022).[4] Table 1 represents some of the primary procedures we implemented in structuring a text. The final product of these text formatting procedures is a vector of sequentially ordered content words.

---

[4] R-package at https://github.com/Reilly-ConceptsCognitionLab/semdistflow

The next step yokes the values of a given database to the ordered content words. The package then derives measures of semantic distance for each word versus its subsequent neighbor.

Table 1. Text stripping, vectorizing, and global formatting operations

Target	Description of Global Action (substitution or omission)
Place Names	Substituted proper noun place names (e.g., 'New York' → 'place') by indexing Wikipedia generated lists of the 100 most populated countries and 100 most visited cities from around the world.
People Names	Substituted common names with (e.g., 'Mohammed' → 'Person') by indexing a list of the 300 most common names from around the world (https://forebears.io/earth/forenames).
Abbreviations	Replaced common abbreviations (e.g., "Dr." → doctor, "Mr." → man).
Contractions	Replaced/extended contractions (e.g., it's → it is).
Letter case	All text converted to lowercase.
Stopwords	Omitted closed class words (e.g., the, a, is) using a custom stopword list (N=1,104 words) modified from the SMART (System for the Mechanical Analysis and Retrieval of Text) stopword list.
Non-alphabetic characters	Omitted all punctuation, symbols, emojis, whitespace, and other non-alphabetic characters.
Numbers	Omitted all cardinal and ordinal numbers.
Morphological Derivatives	Lemmatized the text to transform all words into their corresponding dictionary entries.

Note: In addition to base R, packages used in the various stages of text preparation included TextStem (v0.1.4) (Rinker, 2018a) and Textclean (v0.9.3) (Rinker, 2018b).

2.3 Derivation of a Taxonomic Semantic Space

We created a 15-dimension experiential semantic space characterizing English words (N = 24,771) across a subset of sensorimotor features from the Lancaster Sensorimotor Norms (Lynott et al., 2019) and social-emotional features from the

AffectVec word sentiment norms (Raji & da Melo, 2020). The Lancaster norms reflect crowdsourced salience ratings for 40,000 English words on a 6-point Likert scale. AffectVec reports intensity ratings for 70,000 English words on a 0-1 scale across 239 affective dimensions. We extracted the following sensorimotor dimensions from the Lancaster norms (N = 8): visual, auditory, gustatory, haptic, interoceptive, olfactory, and hand-arm. We extracted the following social-emotional dimensions from AffectVec: excitement, surprise, happiness, fear, anger, contempt, disgust, and sadness. Since these variables reflect different ranges and measurement scales, we z-transformed each individual dimension relative to its own mean and standard deviation. These procedures yielded a vector of z-scores reflecting the salience of each word.

Throughout the remainder of this article, we refer to distance norms generated from this 15-dimension semantic feature space (hereafter SemDist15) as *taxonomic*. This designation is unlikely to meet with universal agreement because taxonomic and thematic semantic relationships tend to be correlated (Kacmajor & Kelleher, 2020). That is, it might be argued that any reference to a semantic space as taxonomic is imprecise or inaccurate because taxonomic knowledge comes with some thematic knowledge. Nevertheless, our rationale for designating SemDist15 as a taxonomic semantic space is that the dimensions comprising this model are feature-based, and feature-based similarity has historically tended to delineate taxonomic relationships (Grossman et al., 2013). We revisit this issue in the results and general discussion.

2.4 Derivation of a Thematic Semantic Space Using Word Embedding

We derived a metric of thematic semantic distance using a well-established word embedding model, GloVe (Global Vectors for Word Representation) (Pennington et al., 2014), trained on documents corresponding to written text[5] within the Corpus of Contemporary American English (CoCA) (Davies, 2009). We first split the corpus into a test set (n=4,273 documents) and a training set (N = 212,737 documents) and omitted all words appearing less than five times. We then trained the model using the text2vec R package (Selivanov, 2020). At a learning rate of 0.05, the embedding model converged after 22 iterations resulting in a 300-dimension vector space spanning 394,115 unique words. We hereafter refer to these embeddings as the Global Vectors of *Written* Contemporary American English (GloWCA).

Throughout the remainder of this article, we refer to bigram distances generated via the GloWCA as *thematic* semantic distance. This designation is likely to incur similar criticisms as denoting SemDist15 as a taxonomic distance measure. Our rationale for

---

[5] In NLP research, a *document* is typically defined as a discrete language sample. Examples of documents include novels, podcast transcripts, blog entries, or transcriptions of spoken language samples. We focused on English text and excluded all documents corresponding to spoken English transcriptions.

describing GloWCA distances as thematic is motivated by face validity. That is, word embeddings are initially generated through co-occurrence statistics similar to the manner in which forks and knives are bound through contextual co-occurrence in kitchens. This distinction of 'what goes together' is the very definition of a thematic semantic relationship.

2.5 Establishing Norms for Thematic and Taxonomic Semantic Distance

Cosine distance values are difficult to interpret in the absence of a known standard. Therefore, our goal in this analysis was to establish norms for taxonomic (using SemDist15) and thematic (using GloWCA) semantic distance for adjacencies between words in continuous written language relative to random word pairings. We first applied the text stripping algorithm described in Section 2.1 to the 4,273 texts in the CoCA test set. In Model 1 (Sequential Model), we randomly selected 250,000 bigrams from the cleaned test set and derived estimates of taxonomic (SemDist15) and thematic (GloWCA) distance for all bigrams. In Model 2 (Random Model), we isolated the initial word within each bigram described in the previous step and then paired each anchor word with another word randomly selected from the test set. In each model, we transformed all semantic distances by taking the inverse and adding one. This transformation constrained each distance to the range 0-2, such that more dissimilar words were associated with greater distances, and more similar words were associated with smaller distances.

2.6 Demonstration Case: *To Build a Fire* (Jack London, 1908)

Our aim in this analysis was to reify the proposed method in a real-world language sample. We analyzed *To Build a Fire* by Jack London (1908), a classic American short story (N = 7,125 words) depicting a man hiking alone through the dark, boreal forest in the Yukon Territory. The man is followed by a native husky or malamute (described in the story as a wolf dog) unbothered by the -75° F temperatures and deep snow. Throughout the story, the dog watches with casual interest as the man attempts to build a fire. His efforts become more and more desperate as hypothermia sets in. The man ultimately freezes to death, and the story ends with the dog wandering off indifferently into the dark forest.

2.6.1 Inter-word Semantic Distance

Our aim was to evaluate how inter-word semantic distance differs between sequential versus randomized text and whether such measures can provide a global index of semantic cohesion (i.e., relatedness of one word to the last). We imported the original text of To Build a Fire into R and executed the cleaning algorithm described in Section 2.1. We retained the ordered version of *To Build a Fire* and created a second version by pairing each word in the story with another word randomly selected from the SemDist15 database. We then derived thematic and taxonomic distances for each bigram and contrasted the random versus the sequential text conditions.

2.6.2 Semantic Flow across *To Build a Fire*

Our aim was to evaluate conceptual variability across the span of a narrative relative to a fixed anchor point. That is, we analyzed semantic distance as a continuous time series reflecting the relationship between new utterances and a block of words initiating the narrative. We reasoned stories are composed of one or more topics. In turn, topics are constrained by constellations of semantically related words. Therefore, punctuated changes in semantic distance are likely to mark topic shifts. Words within a topic (e.g., my day at the zoo) are more semantically related than words spanning different topics (e.g., my day at the zoo vs. my political ideology).

We first cleaned and structured the text of To Build a Fire using the procedures described in Section 2.1. We then isolated the first ten content words in the story to derive a composite semantic vector. This initial block of words served as an anchor point from which we measured semantic distance for every subsequent sentence in the story. We generated composite semantic vectors at the sentence level by averaging vectors for constituent words within each sentence. These procedures generated a time series reflecting semantic dispersion relative to the initial block of words in the story.

3. Results

3.1 Norms for Taxonomic and Thematic Semantic Distance

Table 2 summarizes descriptive statistics for thematic and taxonomic semantic distances for sequential bigrams appearing in naturally ordered text (Sequential Model) versus artificially generated bigrams constructed via random word pairings (Random Model). We normalized cosine distance in all the results to follow by inversing the sign of each value and adding a constant (+1). This procedure yielded values ranging from zero to two where zero indicates no semantic distance (i.e., a word vs. itself) and two reflects the highest possible semantic dissimilarity between two words.
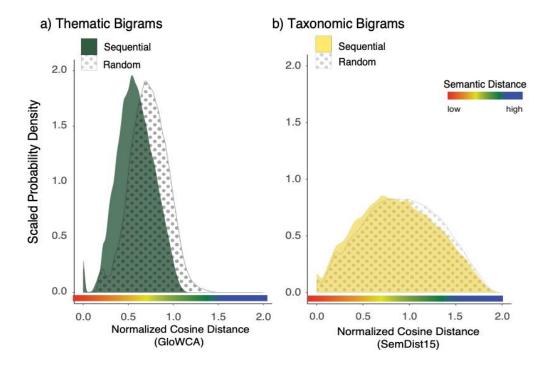
Table 2. Norms for Taxonomic and Thematic Semantic Distance

	Sequential	Random
Thematic (GloWCA)		
Mean (SD)	0.580 (0.204)	0.723 (0.216)
Q1	0.441	0.586
Median	0.575	0.726
Q3	0.723	0.870
Range [Min, Max]	[0, 1.22]	[0, 1.59]
Taxonomic (SemDist15)		
Mean (SD)	0.869 (0.419)	0.921 (0.415)
Q1	0.551	0.610
Median	0.856	0.920
Q3	1.18	1.24
Range [Min, Max]	[0, 1.96]	[0, 1.97]

Note: Distance values reflect the cosine scale transformation

Figure 2 represents distributions of thematic and taxonomic distances for both sequential and random bigrams (N = 250,000 each). We ran two linear regression models to evaluate the relationship between thematic (GloWCA) and taxonomic (SemDist15) distances for the Sequential and Random Models. Taxonomic distances significantly positively predicted thematic distances for the Random Model with a large effect size [Adjusted R^2 = 0.45, F(1, 249998) = 20380, Cohen's f^2 = .81, $p < .001$], indicating that thematic and taxonomic differences capture overlapping semantic information in random bigrams. In the Sequential Model, taxonomic distances significantly positively predicted thematic distances, although with a small effect size [Adjusted R^2 = 0.038, F(1, 249998) = 9978, Cohen's f^2 = .040, $p < .001$]. Thus, although there is overlap in taxonomic and thematic distances, their shared variance is negligible (3.8%) in naturally occurring text [*Spearman's rho* = 0.16, $p < .001$] with a stronger positive correlation in the Random Model [*Spearman's rho* = 0.63, $p < .001$].

Figure 2. Taxonomic and Thematic Distance Distributions

a) Thematic Bigrams

b) Taxonomic Bigrams

Note: The y-axis represents the scaled probability density for a given x-axis value x_n. Probability density is calculated by subtracting the sample mean from x_n and dividing by the standard deviation. The resulting value is then plugged into the Normal probability density function to obtain the probability density for x_n.

3.2 Semantic Distance in a Word of Short Fiction: *To Build a Fire*

Next, we analyze the semantic distances within the sequential and random

versions of To Build a Fire (Table 3). Figure 3 represents the distribution of semantic

distances across all running pairs of content words throughout the story. Taxonomic

semantic distance was moderately higher for randomly generated bigrams relative to

sequential bigrams [Welch-Satterthwaite $t(6381.2) = 12.57$, $p < .001$, Cohen's d = .32

(small to medium effect)]. Random assignment of bigrams resulted in a very large effect

of elevated thematic semantic distance relative to the sequential model [Welch-

Satterthwaite $t(6333.2) = 57.61$, $p < .001$, Cohen's d = 1.41 (very large effect)]. For a

sense of the magnitude of these differences, the effect size for thematic distance in

sequential text was approximately 4.5 times larger than for taxonomic semantic distance.

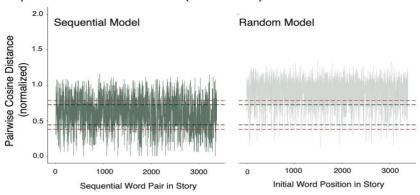Table 3. Taxonomic and Thematic Distances for To Build a Fire

	Condition	N	Mean Dist	SD
Semdist15	Sequential	3134	0.90	.36
(taxonomic)	Randomized	3255	1.01	.36
GloWCA	Sequential	3360	0.62	.23
(thematic)	Randomized	3370	0.91	.18

Note: N= Number of bigrams, Mean Distance reflect cosine values transformed to a 0-2 scale where 0 is the highest possible similarity between two words.
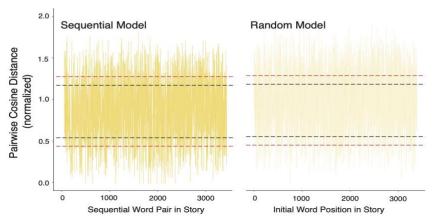
To evaluate whether the findings from the norming study are consistent with the

semantic distances throughout a randomly chosen narrative, we computed mean thematic

and taxonomic distances for sequential and random bigrams in the short story To Build a

Fire. Figure 3 shows the (a) thematic and (b) taxonomic inter-word semantic distances for

each pair of sequential and random bigrams in the narrative.

Figure 3. Thematic and Taxonomic Semantic Distance in To Build a Fire

a) Thematic Distance (GloWCA)

b) Taxonomic Distance (SemDist15)

Note: The red dashed reference lines reflect z-scores corresponding to ± 1.0 based on the norming procedures for adjacencies (i.e., sequential bigrams) described in Section 3.1. The black dashed reference lines reflect the boundaries of Q1 and Q3 reflecting the interquartile ranges derived from the norming study in Section 3.1.

We bandpass-filtered the data pictured in Figure 3 (panel a) to evaluate observed counts of inter-word semantic distances that fell within each of the quantiles established in section 3.1. Table 4 reflects a contingency table of counts for inter-word semantic distances that fell the respective bands of either low distance (< 25th percentile), 'typical'

range (25-75th percentile), or high distance ($>$ 75th percentile). Distributions for thematic

semantic distance ratings in sequential bigrams were independent from the random

bigram distribution, $[\chi^2(2) = 2721, p < .001]$. The distributions for taxonomic semantic

distance ratings in sequential bigrams were also independent from the random bigram

distribution, although the effect was much smaller than in the thematic condition

$[\chi^2(2) = 178, p < .001]$.

Table 4. Contingency Table Representing Counts of Low, Expected, and High Inter-word Semantic Distance in To Build a Fire

Distance Band	Taxonomic		Thematic	
	Sequential	Random	Sequential	Random
Low	817	62	2517	2309
Normal/Medium	2144	546	795	1328
High	1241	3565	665	460

Note: Distance Bands (low, normal, high) correspond to the interquartile range established in Section 3.1 for 250,000 English bigrams. The 'low' band reflects counts of individual observations that fell within the minimum observation to the 25th percentile (Q1). The 'Normal/Medium' distance band reflects Q1 to Q3 (middle 50th percentile). 'High' distance reflects the count of observations within To Build a Fire that exceeded the 75th percentile.

3.3 Sentential Semantic Flow Results: To Build a Fire

This experiment involved analyzing sentence-level changes in word meaning

indexed to a fixed block of ten words initiating the story. Figure 4 illustrates thematic and

taxonomic distance for each sentence in <u>To Build a Fire</u>. We first tested for stationarity and structural continuity (i.e., breakpoint identification) using the 'strucchange' R-package (Zeileis et al., 2002).[6] Both the thematic [automated Dickey Fuller Test Statistic = -4.91, lag order = 7, p = .01] and taxonomic [automated Dickey Fuller Test Statistic = -6.70, lag order = 7, p =.01] time series were stationary. In addition, both time series were characterized by structural changes as confirmed by omnibus tests for thematic [supF = 37.20, p <.001] and taxonomic distances [supF = 27.14, p < .001]. The optimal three-segment partition for the thematic time series had breakpoints at sentence numbers 282 and 344. The optimal three-segment partition for the taxonomic time series had breakpoints at sentence numbers 80 and 205 (see Figure 4).

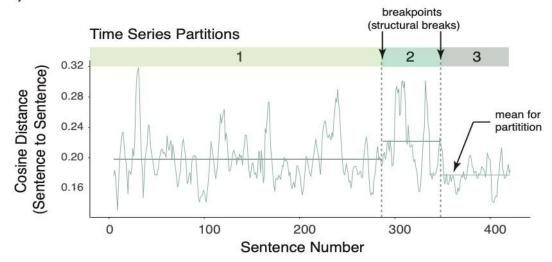Figure 5 illustrates word frequencies across the distinctive partitions (delineated by structural breakpoints) across the taxonomic and thematic distance time series. Frequency distributions qualitatively differed both between (thematic vs. taxonomic) and within (Partitions I, II, III) semantic distance conditions (see Figure 5). In the thematic time series, for example, Partitions I and II differed by an emphasis on environmental descriptors versus words describing interactions between the man and dog. In the

---

[6] Breakpoints correspond to punctuated changes in slope that constitute potential discontinuities or major structural shifts within a continuous time series. Such structural changes effectively partition a time series into subcomponents, each with its own often distinctive mathematical properties (e.g., mean, variance, slope).
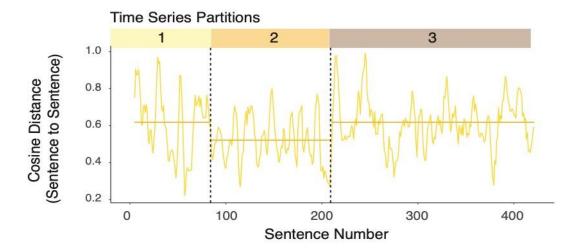
thematic distance condition, breakpoints appear to roughly correspond to the major topic

shifts in the story consistent with the formal assumption of topic modeling that

microevents generate their own unique constellations of words. Although word frequency

distributions also differed within the taxonomic partitions, it was less clear whether these

breakpoints corresponded to topic shifts.

Figure 4. Semantic Flow in <u>To Build a Fire</u>

a) Thematic Semantic Flow: To Build a Fire

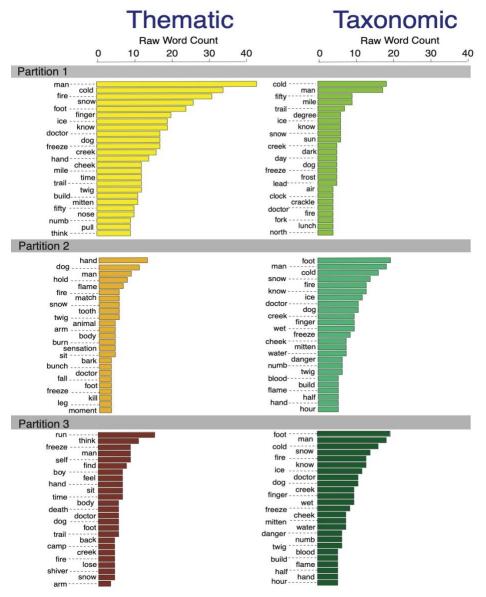# b) Taxonomic Semantic Flow: To Build a Fire

Figure 5. Word Frequency Dispersion Across Semantic Partitions

Note: Barplots represent counts (ranked) of the most 25 frequent words within each partition. For visualization purposes, we omitted the pronouns 'he', 'his', and 'him'.

4. General Discussion

Much remains to be learned about the ways that meaning is conveyed in continuous, natural, language. A deep understanding of the behavior of single words can only take us so far toward the broader aim of deconstructing real world language use. Recent advances in natural language processing have afforded new insights into the ways that words combine at different scales to convey message-level meaning. Semantic distance has played an important role in this transition. Yet, the tools used for generating semantic distances are often complex and their psychological interpretation(s) is not always transparent. In the first part of this article, we discuss thematic and taxonomic semantic systems, a dichotomy driving much recent thought in computer science, as well as cognitive and clinical neuroscience (Brachman, 1983; Kalénine et al., 2012; Schwartz et al., 2011; Zhang et al., 2021). We then review various approaches to measuring semantic distance scaling in complexity from unidimensional constructs (e.g., offensiveness) to high dimensional semantic networks.

Amidst a potentially infinite number of semantic spaces, two dominant approaches to semantic distance are emerging. Word embedding models represent the meaning of each discrete element (e.g., word, phrase, sentence) as a finite vector composed of 50-300 hyperparameters. In contrast, experiential approaches represent word meaning via feature-based salience across a user-defined set of dimensions with ratings reflecting subjective judgments gleaned through crowdsourcing.

We advance a potentially controversial position with respect to the measurement of semantic distance. Namely, we argue that experiential (feature-based) ratings are consistent with taxonomic semantic relationships, whereas word embedding models are more consistent with thematic associations. Just as forks and blenders are commonly encountered together in kitchens, semantically related words are encountered together within circumscribed language topics (e.g., the weather, politics). Knowledge of taxonomic semantic relatedness is also thought to convey some detail about thematic relatedness (Mirman et al., 2017). It could accordingly be argued that semantic spaces are better characterized as existing along a taxonomic-thematic gradient. Here, however, we found that strength of taxonomic-thematic associations for adjacent bigrams in naturally occurring text was quite weak, with only 3.8% shared variance across 250,000 word-pairs. It would therefore appear that the two semantic space models contrasted here (SemDist15 vs. GLoWCA) are measuring different phenomena.

One of our primary aims here was to introduce methods for analyzing inter-word and multiword semantic distances in natural language samples. To this end, we outlined the development of two semantic spaces, each yielding complementary data about changes in word meaning over time. All associated code is freely available for inspection and use within the 'semdistflow' R-package. We encourage researchers to contact us for assistance in applying these algorithms to their own language samples.

One substantive challenge involved in reifying or validating semantic spaces is that any model must be normed against large and representative samples. Section 3.1 involved a norming study where we sampled hundreds of thousands of English bigrams and computed their respective taxonomic and thematic semantic distances from one of the largest commercially available corpora of contemporary written English (i.e., CoCA). These distance norms can serve as the standard against which smaller studies (e.g., picture descriptions by a people with aphasia) might be gauged. In sections 3.2 and 3.3, we applied these norms to assessing word-to-word semantic cohesion and sentence-to-sentence flow in an 'untrained' work of short fiction. We interpret the results of these analyses to follow.

4.1 To Build a Fire: Semantic Distance and Semantic Flow

In the first set of analyses, we analyzed semantic distances for bigrams within *To Build a Fire* using the methods described in Sections 2.1 and 2.2. In a proof-of-concept analysis, we computed mean thematic and taxonomic distances for sequential and random bigrams. Random bigrams showed higher average semantic distances than sequential bigrams. However, this effect (random > sequential) was far stronger for thematically related word pairs, a finding that is consistent with the norming results (Section 3.1). If this inference generalizes, then word-by-word transitions are more predisposed to thematic (e.g., dog-leash) than taxonomic (e.g., dog-wolf) semantic relationships, although both types of relationship are at play.

In a second analysis, we plotted all individual observations within bands corresponding to the Interquartile Ranges for semantic distance established in the norming study (Section 3.1, Figure 2). Several patterns emerged: First, the sequential bigram model (e.g., adjacent words in text) was independent from a random bigram model for both thematic vs taxonomic distance. That is, semantic distances were more likely to fall within the middle 50th percentile established in the norming study for sequential bigrams relative to random bigrams. Again, this effect was far stronger for thematic relative to taxonomic relatedness.

The distinction of 'normal' semantic distance in relation to narrative quality remains unclear. A narrative dominated by low semantic distance could be perceived as repetitive or hyper-focused. In contrast, pathologically high semantic distance could mark narratives as 'word salad' in terms of cohesion and chaining of association. In a companion article (under review), we analyzed tens of thousands of word transitions in the narratives of people with aphasia (Litovsky et al., 2022). Patients showed such non-linearities between semantic distance and conventional text analytics (e.g., mean length of utterance, type-token ratio).

In the final analysis, we explored the 'semantic flow' for each sentence of To Build a Fire relative to the first block of ten content words. We treated semantic distance as a time series and employed statistical methods (e.g., stationarity, structural effects) traditionally used in climate change research and econometrics. The key findings were

that thematic and taxonomic distances yielded qualitatively different time series that produced breakpoints at different parts of the story. These breakpoints partitioned the story into three components, and the distributions of words differed across each of these components. One strong possibility is that thematic semantic distance partitioned the story along major topic transitions. Taxonomic distance also partitioned the story into three discrete sections. However, the nature of these partitions was unclear.

4.2 Potential Applications and Future Directions

The methods proposed here require both fine- and coarse-tuning to achieve this aim. Nevertheless, we envision a variety of applications to conceptualizing language as a continuous time series fluctuating in meaning over time. This includes the following:

a) <u>Causal Modeling</u>: Converting word-to-word level changes in meaning to a numeric time series will potentially facilitate causal modeling of how psycholinguistic (e.g., frequency, word length), psychophysiological (e.g., heart rate, pupil surface area), and neurological (e.g., evoked potentials) variables interact during continuous language perception and production. For example, changes in thematic semantic distance in a running narrative could tax cognitive control which in turn perturbs pupil diameter and heart rate. Yoking numerous simultaneous time series will allow causal modeling during naturalistic language production and comprehension.

b) <u>Implicit Detection of Neurological Disorders</u>: Patterns of impairment in natural language production (e.g., writing samples, spoken language) have proven sensitive to detecting a wide range of neurological and psychiatric disorders (Garrard et al., 2005). Automated screening using implicit language sampling is emerging as a powerful tool for early detection of prodromal dementia (Merkin et al., 2022; Song et al., 2011; Spooner et al., 2020). We know of no current algorithm that considers abnormal semantic distance as a predictor of cognitive impairment. Yet, extensive neuropsychological evidence supports considering such data.

 c) <u>Norming of Developmental Milestones:</u> Little is known about the trajectory of semantic composition throughout early language development. As children learn to narrate written and oral stories, inter-word semantic distance could prove sensitive to gauging maturation of semantic knowledge. Specifically, longitudinal changes in semantic distance during storytelling could yield a sensitive marker of combinatorial semantic abilities.

4.3 Limitations

Language is a rich symbolic modality comprised of many linked processes. Our focus on semantic distance has reduced this complex system to a few simple algorithms looped over large amounts of data. The techniques we have proposed convert language into strings of single words (e.g., DOG-DRINK-WATER). The most obvious limitation of this approach is that language is not simply a string of single words whose meaning

can be linearly decoded. This technique may yield a better understanding of combinatorial semantic processes and cohesion between concepts in narratives. However, it remains to be seen how such information might be integrated with other modalities (e.g., syntax, morphology) to yield a more ecological picture of natural language.

There are several additional limitations and unresolved questions regarding the analyses proposed here. First, our method cannot disambiguate homographs such as *bat,* which could either denote a wooden club or a winged carnivore. This limitation also impacts grammatically ambiguous word forms such as *run.* Our method does not currently employ syntactic parsing methods to disambiguate word meaning using part-of-speech tagging. This decision was motivated by pragmatic considerations related to computational processing demands for people who use our R package. Namely, part-of-speech tagging for longer texts tends be highly resource-intensive. Users can readily modify our R package to incorporate this step. However, the analyses reported here involve matching each word to its most frequent exemplar in a lookup database. This brute-force matching procedure undoubtedly adds error variance, biasing results toward high frequency words.

Another limitation is that the nominal unit of semantic change in continuous language narratives remains undefined. That is, we have proposed methods for modeling word-by-word semantic distance. However, it is possible that people are not particularly sensitive to semantic transitions at the word level and are instead more attuned to the

content of multiword utterances or phrases. This raises potential issues of oversampling

and overfitting that will likely require real-time psycholinguistic approaches (e.g., rapid

serial visual presentation) to ultimately resolve.

4.4 Conclusions

We have proposed a concrete method of analyzing semantic change in language as a

time series. There are many potential applications of this technique, including identifying

semantic impairment using implicit measures, improving automated topic modeling, and

cultivating an understanding of how conceptual combinations arise during early

childhood development. As with any new method, complexities, bugs, and challenges

will emerge. Nevertheless, the potential for modeling language in this way may open a

constellation of forecasting and causal modeling approaches previously reserved for other

disciplines.

Acknowledgement

REFERENCES

Allport, D. A. (1985). Distributed memory, modular subsystems and dysphasia. In S. K. Newman & R. Epstein (Eds.), *Current perspectives in dysphasia* (pp. 207–244). Churchill Livingstone.

Anderson, A. J., Binder, J. R., Fernandino, L., Humphries, C. J., Conant, L. L., Raizada, R. D. S., Lin, F., & Lalor, E. C. (2019). An Integrated Neural Decoder of Linguistic and Experiential Meaning. *Journal of Neuroscience*, *39*(45), 8969–8987. https://doi.org/10.1523/JNEUROSCI.2575-18.2019

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., & Norman, K. A. (2017). Discovering event structure in continuous narrative perception and memory. *Neuron*, *95*(3), 709–721. https://doi.org/10/gbrrws

Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition*, *11*(3), 211–227.

Barsalou, L. W. (1991). Deriving categories to achieve goals. *Psychology of Learning and Motivation*, *27*, 1–64.

Beaty, R. E., & Johnson, D. R. (2021). Automating creativity assessment with SemDis: An open platform for computing semantic distance. *Behavior Research Methods*, *53*(2), 757–780.

Beaty, R. E., Zeitlen, D. C., Baker, B. S., & Kenett, Y. N. (2021). Forward flow and creative thought: Assessing associative cognition and its role in divergent thinking. *Thinking Skills and Creativity*, *41*, 100859.

Beauchamp, M. S., Lee, K. E., Haxby, J. V., & Martin, A. (2002). Parallel visual motion

processing streams for manipulable objects and human movements. *Neuron*,

34(1), 149–159.

Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M.,

& Desai, R. H. (2016). Toward a brain-based componential semantic

representation. *Cognitive Neuropsychology*, 1–45.

https://doi.org/10.1080/02643294.2016.1147426

Binney, R. J., Ashaie, S. A., Zuckerman, B. M., Hung, J., & Reilly, J. (2018).

Frontotemporal stimulation modulates semantically-guided visual search during

confrontation naming: A combined tDCS and eye tracking investigation. *Brain

and Language*, *180*, 14–23.

Brachman, R. J. (1983). What IS-A is and isn't: An analysis of taxonomic links in

semantic networks. *Computer*, *16*(10), 30–36.

https://doi.org/10.1109/MC.1983.1654194

Capitani, E., Laiacona, M., Mahon, B. Z., & Caramazza, A. (2003). What are the facts of

semantic category-specific deficits? A critical review of the clinical evidence.

Cognitive Neuropsychology, *20*(3–6), 213–261.

https://doi.org/10.1080/02643290244000266

Chrysikou, E. G. (2006). When shoes become hammers: Goal-derived categorization

training enhances problem-solving performance. *Journal of Experimental

Psychology: Learning, Memory, and Cognition*, *32*(4), 935.

Cree, G. S., & McRae, K. (2003). Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *Journal of Experimental Psychology: General*, *132*(2), 163–201.

Crutch, S. J., Troche, J., Reilly, J., & Ridgway, G. R. (2013). Abstract conceptual feature ratings: The role of emotion, magnitude, and other cognitive domains in the organization of abstract conceptual knowledge. *Frontiers in Human Neuroscience*, *7*, 186. https://doi.org/10.3389/fnhum.2013.00186

Cutler, A. (1981). Making up materials is a confounded nuisance: Or Will we be able to run any psycholinguistic experiments at all in 1990? *Cognition*, *10*(1–3), 65–70.

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, *14*(2), 159–190. https://doi.org/10/dg6tmh

de Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L., & Theunissen, F. E. (2017). The hierarchical cortical organization of human speech processing. *Journal of Neuroscience*, *37*(27), 6539–6557. https://doi.org/10/gbmfcv

Deniz, F., Nunez-Elizalde, A. O., Huth, A. G., & Gallant, J. L. (2019). The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality. *Journal of Neuroscience*, *39*(39), 7722–7736.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint ArXiv:1810.04805*.

Devlin, J. T., Gonnerman, L. M., Andersen, E. S., & Seidenberg, M. S. (1998). Category-specific semantic deficits in focal and widespread brain damage: A computational account. *Journal of Cognitive Neuroscience*, *10*(1), 77–94.

Dilkina, K., & Lambon Ralph, M. A. (2012). Conceptual Structure within and between Modalities. *Frontiers in Human Neuroscience*, *6*(December), 333. https://doi.org/10.3389/fnhum.2012.00333

Farah, M. J., & McClelland, J. L. (1991). A computational model of semantic memory impairment: Modality specificity and emergent category specificity. *Journal of Experimental Psychology: General*, *120*(4), 339–357.

Fernandino, L., Binder, J. R., Desai, R. H., Pendl, S. L., Humphries, C. J., Gross, W. L., Conant, L. L., & Seidenberg, M. S. (2016). Concept Representation Reflects Multimodal Abstraction: A Framework for Embodied Semantics. *Cerebral Cortex (New York, N.Y.: 1991)*, *26*(5), 2018–2034. https://doi.org/10.1093/cercor/bhv020

Fernandino, L., Tong, J.-Q., Conant, L. L., Humphries, C. J., & Binder, J. R. (2022). Decoding the information structure underlying the neural representation of concepts. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(6), e2108091119. https://doi.org/10.1073/pnas.2108091119

Funnell, E., Hughes, D., & Woodcock, J. (2006). Age of acquisition for naming and knowing: A new hypothesis. *Quarterly Journal of Experimental Psychology*, *59*(2), 268–295.

Garrard, P., Lambon Ralph, M. A., Hodges, J. R., & Patterson, K. (2001). Prototypicality, distinctiveness, and intercorrelation: Analyses of the semantic attributes of living and nonliving concepts. *Cognitive Neuropsychology*, *18*(2), 125–174.

Garrard, P., Maloney, L. M., Hodges, J. R., & Patterson, K. (2005). The effects of very early Alzheimer's disease on the characteristics of writing by a renowned author. *Brain*, *128*(Pt 2), 250–260. https://doi.org/awh341 [pii] 10.1093/brain/awh341 [doi]

Geller, J., Landrigan, J.-F., & Mirman, D. (2019). A pupillometric examination of cognitive control in taxonomic and thematic semantic memory. *Journal of Cognition*, *2*(1). https://doi.org/10/gmb9b2

Gonnerman, L. M., Andersen, E. S., Devlin, J. T., Kempler, D., & Seidenberg, M. S. (1997). Double dissociation of semantic categories in Alzheimer's disease. *Brain and Language*, *57*(2), 254–279.

Grossman, M., McMillan, C., Moore, P., Ding, L., Glosser, G., Work, M., & Gee, J. (2004). What's in a name: Voxel-based morphometric analyses of MRI and naming difficulty in Alzheimer's disease, frontotemporal dementia and corticobasal degeneration. *Brain, 127*(Pt 3), 628–649. https://doi.org/10.1093/brain/awh075

Grossman, M., Peelle, J. E., Smith, E. E., McMillan, C. T., Cook, P., Powers, J., Dreyfuss, M., Bonner, M. F., Richmond, L., Boller, A., Camp, E., & Burkholder, L. (2013). Category-specific semantic memory: Converging evidence from bold fMRI and Alzheimer's disease. *Neuroimage*, *68*, 263–274. https://doi.org/10.1016/j.neuroimage.2012.11.057

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

Hartung, F., Kenett, Y. N., Cardillo, E. R., Humphries, S., Klooster, N., & Chatterjee, A. (2020). Context matters: Novel metaphors in supportive and non-supportive contexts. *NeuroImage*, *212*, 116645.

Hillis, A. E., & Caramazza, A. (1991). Category-specific naming and comprehension impairment: A double dissociation. *Brain*, *114, Pt 5*, 2081–2094.

Hills, T. T., & Kenett, Y. N. (2021). Networks of the Mind: How Can Network Science Elucidate Our Understanding of Cognition? *Topics in Cognitive Science*.

Hills, T. T., Todd, P. M., & Jones, M. N. (2015). Foraging in Semantic Fields: How We Search Through Memory. *Topics in Cognitive Science*, n/a-n/a. https://doi.org/10.1111/tops.12151

Hodges, J. R., Patterson, K., Graham, N., & Dawson, K. (1996). Naming and knowing in dementia of Alzheimer's type. *Brain and Language*, *54*(2), 302–325. https://doi.org/S0093-934X(96)90077-2 [pii] 10.1006/brln.1996.0077 [doi]

Hoffman, P., McClelland, J. L., & Lambon Ralph, M. A. (2018). Concepts, control, and context: A connectionist account of normal and disordered semantic cognition. *Psychological Review*, *125*(3), 293. https://doi.org/10/gdjrq9

Hoffman, P., Rogers, T. T., & Lambon Ralph, M. A. (2011). Semantic diversity accounts for the "missing" word frequency effect in stroke aphasia: Insights using a novel method to quantify contextual variability in meaning. *Journal of Cognitive Neuroscience*, *23*(9), 2432–2446.

Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458. https://doi.org/10/bfqp

Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, *76*(6), 1210–1224.

Jackendoff, R. (1987). On Beyond Zebra: The relation of linguistic and visual information. *Cognition*, *26*(2), 89–114. https://doi.org/10.1016/0010-0277(87)90026-6

Jain, S., & Huth, A. G. (2018). Incorporating context into language encoding models for fMRI. *BioRxiv*, 327601.

Kacmajor, M., & Kelleher, J. D. (2020). Capturing and measuring thematic relatedness. *Language Resources and Evaluation*, *54*(3), 645–682. https://doi.org/10/gn655n

Kalénine, S., Mirman, D., & Buxbaum, L. J. (2012). A combination of thematic and similarity-based semantic processes confers resistance to deficit following left hemisphere stroke. *Frontiers in Human Neuroscience*, *6*(May), 106. https://doi.org/10.3389/fnhum.2012.00106

Kalénine, S., Peyrin, C., Pichat, C., Segebarth, C., Bonthoux, F., & Baciu, M. (2009). The sensory-motor specificity of taxonomic and thematic conceptual relations: A behavioral and fMRI study. *NeuroImage*, *44*(3), 1152–1162. https://doi.org/10.1016/j.neuroimage.2008.09.043

Kim, J. S., Elli, G. V., & Bedny, M. (2019). Knowledge of animal appearance among sighted and blind adults. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(23), 11213–11222. https://doi.org/10.1073/pnas.1900952116

Kousta, S.-T. T., Vigliocco, G., Vinson, D. P., Andrews, M., & Del Campo, E. (2011). The representation of abstract words: Why emotion matters. *Journal of Experimental Psychology General*, *140*(1), 14–34. https://doi.org/2010-26153-001 [pii] 10.1037/a0021446 [doi]

Kumar, A. A. (2021). Semantic memory: A review of methods, models, and current challenges. *Psychonomic Bulletin & Review*, *28*(1), 40–80.

Kumar, A. A., Steyvers, M., & Balota, D. A. (2021). A Critical Review of Network-Based and Distributional Approaches to Semantic Memory Structure and Processes. *Topics in Cognitive Science*, *0*(n/a). https://doi.org/10.1111/tops.12548

Landauer, T. K., & Dumais, S. T. (1997). Solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

Litovsky, C., Finley, A., Zuckerman, B., Sayers, M., Schoenhard, J., Kennett, Y., & Reilly, J. (under review, 2022). Semantic flow and its relation to controlled lexical-semantic retrieval deficits in aphasia.

London, J. (1908). *To Build a Fire*. The Century.

Lupker, S. J. (1979). The semantic nature of response competition in the picture-word interference task. *Memory & Cognition*, *7*(6), 485–495.

Maki, W. S., & Buchanan, E. (2008). Latent structure in measures of associative, semantic, and thematic knowledge. *Psychonomic Bulletin & Review*, *15*(3), 598–603. https://doi.org/10/b8kgbb

Malt, B. C. (2020). Words, thoughts, and brains. *Cognitive Neuropsychology*, *37*(5–6), 241–253.

Malt, B. C., Ameel, E., Gennari, S., Imai, M., Saji, N., & Majid, A. (2011). Do words reveal concepts? *Proceedings of the Annual Meeting of the Cognitive Science Society*, *33*(33).

Malt, B. C., & Majid, A. (2013). How thought is mapped into words. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(6), 583–597. https://doi.org/10.1002/wcs.1251

Mandera, P., Keuleers, E., & Brysbaert, M. (2015). How useful are corpus-based

methods for extrapolating psycholinguistic variables? *Quarterly Journal of*

Experimental Psychology, *68*(8), 1623–1642.

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in

psycholinguistic tasks with models of semantic similarity based on prediction and

counting: A review and empirical validation. *Journal of Memory and Language*,

92, 57–78.

Martin, A. (2015). GRAPES—Grounding representations in action, perception, and

emotion systems: How object properties and categories are represented in the

human brain. *Psychonomic Bulletin & Review*, 1–12.

https://doi.org/10.3758/s13423-015-0842-3

McNorgan, C., Reid, J., & McRae, K. (2011). Integrating conceptual knowledge within

and across representational modalities. *Cognition*, *118*(2), 211–233.

https://doi.org/S0010-0277(10)00254-4 [pii] 10.1016/j.cognition.2010.10.017

[doi]

McRae, K., Cree, G. S., Westmacott, R., & Sa, V. R. D. (1999). Further evidence for

feature correlations in semantic memory. *Canadian Journal of Experimental*

Psychology/Revue Canadienne de Psychologie Expérimentale, *53*(4), 360.

Merkin, A., Krishnamurthi, R., & Medvedev, O. N. (2022). Machine learning, artificial

intelligence and the prediction of dementia. *Current Opinion in Psychiatry*, *35*(2),

123–129. https://doi.org/10.1097/YCO.0000000000000768

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 3111–3119.

Mirman, D., Landrigan, J.-F., & Britt, A. E. (2017). Taxonomic and thematic semantic systems. *Psychological Bulletin*, *143*(5), 499. https://doi.org/10/gkbqjp

Mota, N. B., Copelli, M., & Ribeiro, S. (2017). Thought disorder measured as random speech structure classifies negative symptoms and schizophrenia diagnosis 6 months in advance. *Npj Schizophrenia*, *3*(1), 1–10. https://doi.org/10.1038/s41537-017-0019-3

Mota, N. B., Vasconcelos, N. A., Lemos, N., Pieretti, A. C., Kinouchi, O., Cecchi, G. A., Copelli, M., & Ribeiro, S. (2012). Speech graphs provide a quantitative measure of thought disorder in psychosis. *PloS One*, *7*(4), e34928. https://doi.org/10.1371/journal.pone.0034928

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *Neuroimage*, *56*(2), 400–410.

Nastase, S. A., Goldstein, A., & Hasson, U. (2020). Keep it real: Rethinking the primacy of experimental control in cognitive neuroscience. *NeuroImage*, *222*, 117254.

Patterson, K. (2007). The reign of typicality in semantic memory. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *362*(1481), 813–821. https://doi.org/1MW426N6VV32M223 [pii] 10.1098/rstb.2007.2090 [doi]

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10/gfshwg

Peters, M. E., Ammar, W., Bhagavatula, C., & Power, R. (2017). Semi-supervised sequence tagging with bidirectional language models. *ArXiv Preprint ArXiv:1705.00108*.

Pexman, P. M., Heard, A., Lloyd, E., & Yap, M. J. (2016). The Calgary semantic decision project: Concrete/abstract decision data for 10,000 English words. *Behavior Research Methods*, 1–11. https://doi.org/10.3758/s13428-016-0720-6

Popham, S. F., Huth, A. G., Bilenko, N. Y., Deniz, F., Gao, J. S., Nunez-Elizalde, A. O., & Gallant, J. L. (2021a). Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nature Neuroscience*, *24*(11), 1628–1636.

Popham, S. F., Huth, A. G., Bilenko, N. Y., Deniz, F., Gao, J. S., Nunez-Elizalde, A. O., & Gallant, J. L. (2021b). Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nature Neuroscience*, *24*(11), 1628–1636. https://doi.org/10/gm9zp2

Prabhakaran, R., Green, A. E., & Gray, J. R. (2014). Thin slices of creativity: Using single-word utterances to assess creative cognition. *Behavior Research Methods*, *46*(3), 641–659. https://doi.org/10.3758/s13428-013-0401-7

Price, A. R., Bonner, M. F., Peelle, J. E., & Grossman, M. (2015). Converging evidence for the neuroanatomic basis of combinatorial semantics in the angular gyrus. *The Journal of Neuroscience*, *35*(7), 3276–3284. https://doi.org/10.1523/JNEUROSCI.3446-14.2015

Raji, S., & da Melo, G. (2020). What sparks joy: The AffectVec emotion database. *Proceedings of the Web Conference, ACM*.

Reilly, J., Finley, A. M., Kelly, A., Zuckerman, B., & Flurie, M. (2021). Olfactory language and semantic processing in anosmia: A neuropsychological case control study. *Neurocase*, 1–11. https://doi.org/10/gh6bdf

Reilly, J., Peelle, J. E., Garcia, A., & Crutch, S. J. (2016). Linking somatic and symbolic representation in semantic memory: The dynamic multilevel reactivation framework. *Psychonomic Bulletin & Review*, *23*(4), 1002–1014. https://doi.org/10.3758/s13423-015-0824-5

Rinker, T. W. (2018a). *Textstem: Tools for stemming and lammatizing* (0.1.4) [Computer software].

Rinker, T. W. (2018b). *Textclean* (0.9.3) [Computer software]. https://CRAN.R-project.org/package=textclean

Rogers, T. T., Lambon Ralph, M. A., Garrard, P., Bozeat, S., McClelland, J. L., Hodges, J. R., & Patterson, K. (2004). Structure and deterioration of semantic memory: A Neuropsychological and computational investigation. *Psychological Review*, *111*(1), 205–235.

Rosch, E. H. (1973). Natural categories. *Cognitive Psychology*, *4*(3), 328–350. https://doi.org/10.1016/0010-0285(73)90017-0

Sacchett, C., & Humphreys, G. W. (1992). Calling a squirrel a squirrel but a canoe a wigwam: A category-specific deficit for artefactual objects and body parts. *Cognitive Neuropsychology*, *9*(1), 73–86.

Sadeghi, Z., Mcclelland, J. L., & Hoffman, P. (2014). You shall know an object by the company it keeps: An investigation of semantic representations derived from object co-occurrence in visual scenes. *Neuropsychologia*, *2012*, 1–10. https://doi.org/10.1016/j.neuropsychologia.2014.08.031

Sanz, C., Pallavicini, C., Carrillo, F., Zamberlan, F., Sigman, M., Mota, N., Copelli, M., Ribeiro, S., Nutt, D., & Carhart-Harris, R. (2021). The entropic tongue: Disorganization of natural language under LSD. *Consciousness and Cognition*, *87*, 103070. https://doi.org/10.1016/j.concog.2020.103070

Schwartz, M. F., Kimberg, D. Y., Walker, G. M., Brecher, A., Faseyitan, O. K., Dell, G. S., Mirman, D., & Coslett, H. B. (2011). Neuroanatomical dissociation for taxonomic and thematic knowledge in the human brain. *Proceedings of the National Academy of Sciences*, *108*(20), 8520–8524.

Selivanov, D. (2020). *text2vec: Modern text mining framework for R* (0.6) [Computer software]. https://cran.r-project.org/web/packages/text2vec/index.html

Simony, E., Honey, C. J., Chen, J., Lositsky, O., Yeshurun, Y., Wiesel, A., & Hasson, U. (2016). Dynamic reconfiguration of the default mode network during narrative comprehension. *Nature Communications*, *7*(1), 1–13. https://doi.org/10/f8xpfb

Song, X., Mitnitski, A., & Rockwood, K. (2011). Nontraditional risk factors combine to predict Alzheimer disease and dementia. *Neurology*, *77*(3), 227–234. https://doi.org/10.1212/WNL.0b013e318225c6bc

Spooner, A., Chen, E., Sowmya, A., Sachdev, P., Kochan, N. A., Trollor, J., & Brodaty, H. (2020). A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Scientific Reports*, *10*(1), 1–10. https://doi.org/10.1038/s41598-020-77220-w

Troche, J., Crutch, S. J., & Reilly, J. (2017). Defining a conceptual topography of word concreteness: Clustering properties of emotion, sensation, and magnitude among 750 English words. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.01787

Trumpp, N. M., Kliese, D., Hoenig, K., Haarmeier, T., & Kiefer, M. (2013). Losing the sound of concepts: Damage to auditory association cortex impairs the processing of sound-related concepts. *Cortex*, *49*(2), 474–486.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x

Warrington, E. K. (1975). The selective impairment of semantic memory. *Quarterly Journal of Experimental Psychology*, *27*(4), 635–657.

Westerlund, M., & Pylkkänen, L. (2014). The role of the left anterior temporal lobe in semantic composition vs. Semantic memory. *Neuropsychologia*, *57*, 59–70. https://doi.org/10.1016/j.neuropsychologia.2014.03.001

Woollams, A. M., Cooper-Pye, E., Hodges, J. R., & Patterson, K. (2008). Anomia: A doubly typical signature of semantic dementia. *Neuropsychologia*, *46*(10), 2503–2514.

Zeileis, A., Leisch, F., Hornik, K., & Kleiber, C. (2002). strucchange: An R package for testing for structural change in linear regression models. *Journal of Statistical Software*, *7*, 1–38. https://doi.org/10.18637/jss.v007.i02

Zhang, M., Varga, D., Wang, X., Krieger-Redwood, K., Gouws, A., Smallwood, J., & Jefferies, E. (2021). Knowing what you need to know in advance: The neural processes underpinning flexible semantic retrieval of thematic and taxonomic relations. *NeuroImage*, *224*, 117405. https://doi.org/10.1016/j.neuroimage.2020.117405

Zuckerman, B., Reilly, J., Litovsky, C., & Finley, A. (2021). *semdistflow: Strip, format, vectorize text to sequential bigram semantic distance* (1.0) [Computer software