



Published in final edited form as:

J Exp Psychol Gen. 2023 September ; 152(9): 2578–2590. doi:10.1037/xge0001389.

Bigram Semantic Distance as an Index of Continuous Semantic Flow in Natural Language: Theory, Tools, and Applications

Jamie Reilly^{1,2}, Ann Marie Finley^{1,2}, Celia P. Litovsky^{1,2}, Yoed N. Kenett³

¹Eleanor M. Saffran Center for Cognitive Neuroscience

²Department of Communication Sciences and Disorders, Temple University, Philadelphia, Pennsylvania USA

³Faculty of Faculty of Data and Decision Sciences, Technion Israel Institute of Technology, Haifa, Israel

Abstract

Much of our understanding of word meaning has been informed through studies of single words. High dimensional semantic space models have recently proven instrumental in elucidating connections *between* words. Here we show how bigram semantic distance can yield novel insights into conceptual cohesion and topic flow when computed over continuous language samples. For example, ‘*Cats drink milk*’ is comprised of an ordered vector of bigrams (cat-drink, drink-milk). Each of these bigrams has a unique semantic distance. These distances in turn may provide a metric of dispersion or the flow of concepts as language unfolds. We offer an R package (‘semdistflow’) that transforms any user-specified language transcript into a vector of ordered bigrams, appending two metrics of semantic distance to each pair. We validated these distance metrics on a continuous stream of simulated verbal fluency data assigning predicted switch markers between alternating semantic clusters (animals, musical instruments, fruit). We then generated bigram distance norms on a large sample of text and demonstrated applications of the technique to a classic work of short fiction, *To Build a Fire* (Jack London, 1908). In one application, we showed that bigrams spanning sentence boundaries are punctuated by jumps in semantic distance. We discuss the promise of this technique for characterizing semantic processing in real world narratives and for bridging findings at the single word level with macroscale discourse analyses.

Keywords

Semantic Distance; Semantic Memory; Language; Narrative; Natural Language Processing

Address Correspondence to: Jamie Reilly, Ph.D., Temple University, 1701 N. 13th Street, Philadelphia, PA 19122, reillyj@temple.edu.

CRediT author statement

Jamie Reilly: Conceptualization, Methodology, Software, Writing; **Celia Litovsky:** Methodology, Writing- Reviewing and Editing;

Ann Marie Finley: Methodology, Writing- Reviewing and Editing; **Yoed Kenett:** Methodology, Writing- Reviewing and Editing

Prior Dissemination

Portions of this work appear/ed as a preprint on PsyArXiv (<https://psyarxiv.com/6fuhv>) and were presented at the annual conference of the Academy of Aphasia (2022)

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 et al., 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 & Pykkänen, 2014; Marelli, Gagné, & Spalding, 2017).

Techniques leveraged from natural language processing 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; Johnson et al., 2022; Kumar, 2021; Kumar et al., 2022; 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 (e.g., patterns of co-occurrence) about language by indexing vast numbers of words 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) to construct high-dimensional models of word meaning. Such models have fueled numerous recent advances across a wide range of psychological and linguistic sciences (Beaty et al., 2021; Beaty & Johnson, 2021; Gray et al., 2019; Hills & Kenett, 2022; Johnson et al., 2022; Kenett, 2018, 2019; Kenett et al., 2017; Kumar et al., 2020; Olson et al., 2021), as well as within computational, clinical, and cognitive neuroscience (Anderson et al., 2019; Fernandino et al., 2016, 2022, Fu et al., 2022; Kenett & Faust, 2019).

1.1 What is semantic distance?

Semantic distance reflects the similarity (or dissimilarity) between two or more concepts distributed across an n -dimensional space, typically derived from analyzing large corpora of texts (Günther et al., 2019). There is no upward limit to the number of potential dimensions that comprise a semantic space. A researcher with an interest in arousal of curse words, for example, might quantify differences in the arousal ratings of curse words vs. neutral words (e.g., Reilly et al., 2020). In this simple example, *arousal* constitutes a one-dimensional semantic space.

Semantic distance is a relative rather than absolute construct. That is, semantic distance can only be interpreted relative to the unique semantic space used to derive that measure. Cognitive (neuro)scientists are typically interested in constructing semantic spaces that are

psychologically and/or neurobiologically plausible (Binder et al., 2016; Crutch et al., 2013; Reilly et al., 2021; Sacchett & Humphreys, 1992). An existential challenge for semantic space models is that the true dimensionality of human semantic memory is latent. My semantic network is qualitatively different than yours, and any viable model of semantic memory must have the flexibility to accommodate these differences (Kumar, 2021).

Two broad classes of semantic space models have risen to prominence over the past decade. Experiential models classify concepts along subjectively experienced dimensions (e.g., color, shape), such that words with similar characteristics are more semantically linked. In contrast, word embedding models are predicated upon the idea that in natural language, words that occur together are likely to be semantically related. Both models approach the challenge of specifying dimensionality in fundamentally different ways. In the following sections, we describe both types of models and how they capture meaning through multidimensional vectors.

1.2 Experiential semantic models

A core assumption underlying all semantic space models is that word meaning can be decomposed into numerous dimensions or features (for early iterations of feature-based approaches see Cree & McRae, 2003; Rosch, 1973). In *experiential* models, raters are typically asked to explicitly rate the salience of many target words across numerous orthogonalized dimensions. For example, a researcher might ask people to rate the salience of artichokes on color, visual imagery, and aggression. These ratings rely on our own subjective experience and as such, have been termed *experiential* (Binder et al., 2016; see also Wingfield, & Connell, 2022). Binder and colleagues (2016) have proposed perhaps the most extensive experiential model to date, characterizing words along 65 sensorimotor, affective, and interoceptive dimensions. In an earlier model known as the Abstract Conceptual Feature space, Crutch et al. characterized English words along twelve dimensions (Crutch et al., 2013; Reilly et al., 2016; Troche et al., 2017). Both the Binder et al (2016) and Crutch et al (2013) experiential models tend to produce clusters that mirror classic Linnaean taxonomies. That is, similarities in color, sound, valence, danger, body morphology, and other traits form constellations of intercorrelated features that bound natural categories (see also Cree & McRae, 2003; Garrard et al., 2001; McRae et al., 1999; Rogers et al., 2004). In an experiential semantic space, dogs and wolves have similar vector representations since they highly overlap in shape, color, sound, and other characteristics. However, experiential models do not typically account for contextual or thematic relatedness predicated upon co-occurrence. For this type of semantic relation, we turn to word embedding models, which quantify regularities in the linguistic contexts in which words co-occur (i.e., are embedded).

1.3 Word embeddings, context, and co-occurrence

In contrast to experiential semantic models, embedding models capture semantic similarity based on shared environments. For example, dogs, collars, bones, frisbees, and leashes are all semantically bound via shared contexts. We learn through repeated exposure that the conditional probability of encountering a collar in the context of a dog is high. Common 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-3 (Radford et al., 2018) attempt to capture this co-occurrence in language by modeling abstract vector representations such that words that co-occur more frequently have more similar vector representations.¹ In embedding models, dogs and collars are thus closely linked (with similar vector representations), whereas dogs and platypuses, which do not frequently co-occur, have more dissimilar vector representations. Yet dogs and platypuses share many features (e.g., both are mammals), and thus are likely to share similar *experiential* semantic vector representations.

Word embeddings are typically derived through machine learning algorithms that require no subjective human judgments. Such models are typically trained on vast language corpora (i.e., collections of structured text such as books, newscasts, transcribed podcasts, or Twitter feeds). The outcome of this process is that each discrete token (e.g., word, *n*-gram, phrase) in a training corpus is characterized by a multidimensional array. However, unlike the labeled dimensions that comprise experiential semantic models, the dimensions derived from word embeddings are abstract mathematical constructs.

1.4 Inter-word Semantic Distance: A Continuous Bigram Model

Semantic space models have featured prominently in many recent language and neuroimaging studies (Bonner & Epstein, 2022; Binder et al., 2016; Crutch et al., 2013; Marelli, Gagné, & Spalding, 2017; Reilly et al., 2021; Wingfield and Connell, 2022). However, to our knowledge these models have not been applied to evaluate the flow of semantic information in continuous language. Consider, for example, the following sentence:

The quick brown fox jumped over the lazy dog.—After omitting closed class words (e.g., *the*), this sentence is composed of five lemmatized bigrams, including: 1) quick-brown; 2) brown-fox; 3) fox-jump; 4) jump-lazy; 5) lazy-dog. When a semantic distance is assigned to each bigram, a vector of distances emerges. This vector is technically a continuous time series reflecting word-by-word conceptual shifts over the course of any language sample.

Many previous studies investigating semantic distance have employed ‘continuous bag of words’ (or CBOW) approaches that either abandon order information or alternatively analyze semantic distance for paradigms where computing continuous bigram distance would be senseless (e.g., word association). However, under a sequential bigram model, order is a critical factor reflecting dispersion of concepts over time. For example, a person experiencing severe mania or delirium might show markedly high semantic distance perceived by a listener as incoherent. In contrast, low variability in semantic distance could mark a narrowly focused, boring, or repetitive story. We recently reported evidence of such effects using a ‘bag of bigrams’ approach to spoken narratives produced by people with aphasia relative to age-matched controls (Litovsky et al., 2022). People with aphasia showed significantly lower bigram semantic distances relative to controls, and semantic

¹Many contemporary transformer models (e.g., BERT, GPT-3) use a combination of lexical co-occurrence and deep learning to extract semantic vectors. Thus, references to embedding models as ‘co-occurrence models’ is a misnomer.

distances correlated strongly with offline neuropsychological measures of semantic memory functioning.

Our aims in the current study were to describe the development and implementation of a freely available R package designed to read, clean, tokenize, and append two novel metrics of semantic distance (i.e., experiential vs. embedding) to any continuous language sample. We derived norms for each of these semantic spaces and conducted a validation study demonstrating how jumps in semantic distance can mark cluster boundaries in a continuous stream of simulated category fluency data (i.e., blocks of musical instruments, fruits, animals). Finally, we demonstrate several novel applications of this continuous bigram model to a renowned work of short fiction, *To Build a Fire* (Jack London, 1908).

2 Methods

2.1 Overview

We first derived two novel semantic spaces (experiential and embedding) that provide the foundation for an open-source R package, titled ‘semdistflow’. This R package reads, formats, and then transforms any language transcript into a running vector of pairwise semantic distances. We derived distance norms for each semantic space, validated the spaces on a continuous vector of alternating semantic categories, and conducted a series of simulations on a work of short fiction. All scripts and data used here are available for download and use at <https://osf.io/ryhfj/>.

2.2 Derivation of a Feature-based Semantic Space (SemDist15)

We created a 15-dimension experiential semantic space characterizing English words 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: 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 *experiential*.

2.3 Derivation of an Embedding Semantic Space (GloWCA)

We derived an embedding space using a well-established model, GloVe (Global Vectors for Word Representation; Pennington et al., 2014), trained on documents corresponding to written text² 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 fewer 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 trimmed this large data file (2.2 Gb) by isolating only those entries with a corresponding lemma listed in the Subtlex-US word frequency norms (Brysbaert and New, 2009), resulting in 60,384 words, each characterized across 300 hyperparameters. 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 distance norms generated from GloWCA as *embedding*.

2.4 Text Cleaning Algorithm

To apply vector representations to each word in a language sample, we first isolated all content words and reduced them to their respective lemma forms. This cleaning algorithm involved multiple steps (see Table 1) followed by lemmatization of the remaining content words. Readers are invited to inspect all individual commands by visiting the R-package vignette for ‘semdistflow’ (Zuckerman, Reilly, Litovsky, & Finley, 2022).³ Table 1 represents some of the primary procedures we implemented in this cleaning algorithm. The final product of these text cleaning and formatting procedures is a vector of sequential bigrams (i.e., each word paired with the next content word in the language sample) to which we append our experiential semantic space (SemDist15) and embedding semantic space (GloWCA) distance values.

2.5 Establishing Norms for Semantic Distance

In this analysis, we calculated semantic distances in a large sample of naturally occurring sequential bigrams relative to randomly paired words to establish norms for semantic distance in natural language. We calculated semantic distance via the cosine similarity measure. Cosine is a scalar value that measures the similarity of two vectors’ angles. For example, to calculate the experiential semantic distance of “quick” and “brown,” we calculated the cosine similarity of the 15-dimensional vectors of “quick” and “brown” represented in SemDist15. We transformed all semantic distances by subtracting the cosine similarity from 1. This procedure constrained each semantic distance to a range between zero and two, such that more dissimilar words were associated with greater distances, and more similar words were associated with smaller distances.

After processing the 4,273 texts in the CoCA test set using our text cleaning algorithm, we extracted 250,000 sequential bigrams to establish norms for semantic distances in naturally occurring language. In this Sequential Model, we derived estimates of experiential (SemDist15) and embedding (GloWCA) semantic distances. As a comparison (Random Model), we yoked the initial word of each bigram to another randomly selected word in the test set, yielding non-sequential bigrams.

²In 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.

³R-package at <https://github.com/Reilly-ConceptsCognitionLab/semdistflow>

2.6.1 Validation using simulated category fluency data

Semantic fluency tasks typically involve producing as many exemplars of a given category (e.g., animals) as possible over a fixed interval (Christensen & Kenett, 2021). People commonly employ foraging and other search strategies such as clustering and switching when producing a string of exemplars (Binney et al., 2018; Troyer et al., 1998; Troyer, 2000; Ovando-Tellez et al., 2022). For example, a category such as animals might evoke a spontaneous clustering and switching strategy involving house pets switching to jungle animals, marine animals, etc. Successful verbal fluency, therefore, requires both effective semantic search within categories, as well as fluid executive functioning as needed to flexibly disengage when a particular category has been exhausted (Ovando-Tellez et al., 2022). Verbal fluency has accordingly emerged as one of the most common metrics of executive and semantic processing employed in clinical neuropsychology.

The overarching structure of semantic category fluency in terms of alternating blocks of semantic clusters may offer a unique opportunity to validate our proposed continuous bigram model. We reasoned that semantic distance within semantic clusters will be low (e.g., dog, cat, hamster) relative to semantic distances for bigrams crossing switch boundaries (e.g., ‘dog – cat - **hamster** | **saxophone** -piano -trumpet...’). Thus, large jumps in semantic distance can potentially delineate switches in a continuous stream of category fluency data (for related work on segmenting fluency data using word embedding models see Alacam et al., 2022; Lundin et al., 2022).

We first generated a vector of 7500 continuous words composed of alternating 10-word blocks of animals, musical instruments, and fruits/vegetables randomly sampled with replacement from fixed lists (see OSF for lists). This vector was therefore, composed of 750 switches and 6750 within cluster exemplars, providing a fixed reference for exactly where switches occur (i.e., every tenth word). We applied the `distme()` function to the unlemmatized ordered word list, generating pairwise semantic distances for every running bigram. We then scaled (z-scored) the resultant distributions and recoded each running word pair as potentially either a within category cluster (coded as 0) or as a predicted switch between categories (coded as 1) using a threshold of $z > 1$. We examined concordance between the actual distribution of switches (every tenth word) versus the predicted distribution of switches (marked by z-score distance jumps) using a variety of signal detection metrics from the ‘verification’ package (NCAR Research Applications Library, 2015) of R.

2.6.2 To Build a Fire (Jack London, 1908)—We hypothesized that sequential bigrams as naturally occurring within structured text are more semantically related than random bigrams. Semantic distance should on average be lower for sequential bigrams relative to randomly constructed bigrams (i.e., each word in the story paired with a random word in another database). We tested this hypothesis in the context of *To Build a Fire*, a novella published by the American author, Jack London, in 1908. This famous story ($N = 7,125$ words) depicts a man hiking alone through the boreal forest of the Yukon Territory. The man is followed by a native dog (described as a *wolf dog*) unfazed by the deep snow and cold (-75° F). The dog casually follows the man from a distance as his attempts at

self-preservation by building a fire become increasingly more desperate. The man ultimately succumbs to hypothermia, and the story ends with the dog wandering off indifferently into the forest.

We first imported the original text into R and executed the cleaning algorithm described in Section 2.4. We computed semantic distances for the sequential bigrams from the original story. We then created a second random bigram vector consisting of each lemma of *To Build a Fire* in its original order randomly paired with an entry from the SemDist15 database. We derived experiential and embedding semantic distances for bigrams in both the sequential and random conditions.

2.6.3 Semantic Distance across Sentence Boundaries in *To Build a Fire*—

Here we tested the hypothesis that sentences constitute micro-topics with higher semantic relatedness (low semantic distance) within a sentence than between adjacent sentences. We tested this prediction by contrasting semantic distances (embedding and experiential) for sequential bigrams within sentences (within condition) relative to bigrams that broke across a sentence boundary (switch condition). We first cleaned the text of *To Build a Fire* using the procedures described in Section 2.4. We then coded each bigram as either within (non-adjacent to a period) or as a switch trial. Switch bigrams consisted of the final word of one sentence paired with the initial word of the following sentence.

3. Results

3.1 Norms for Bigram Semantic Distance

Tables 2 and 3 summarize descriptive statistics for experiential and embedding semantic distances for sequential bigrams in naturally ordered text (Sequential Model) versus artificially generated bigrams constructed via random word pairings (Random Model). We assigned categorical ranges ('low', 'medium', 'high') by referencing the interquartile range for each condition. The first and fourth quartile constitute 'low' and 'high' distance, whereas the middle quartiles (Q1 to Q3) reflect medium or expected distances.

Figure 1 reflects density plots for both semantic spaces ($N = 250,000$ each). The distribution of bigram semantic distances generated by the GloWCA embedding model was tightly clustered and leptokurtic relative to the experiential semantic distance distribution whose variance was over four times higher.

3.2 Category fluency validation

We simulated a continuous stream of category fluency data consisting of 7500 words with blocks of alternating 10-word clusters in a fixed order (i.e., animals, musical instruments, fruits/vegetables). This 'actual' time series consisted of 6750 words within clusters and 749 switches (i.e., there is no semantic distance or switch for the final word in the series). We coded all bigrams as either within cluster ($z < 1$) or as constituting a switch point between clusters ($z \geq 1$) using their scaled cosine distance. Table 4 and Figure 2 represent comparisons of the predicted distribution of binary events (switches and clusters) to the actual distribution of events. Overall classification accuracy was similar between the

experiential (semDist15, 91.5%) and embedding spaces (GloWCA, 90.9%) with a medium tetrachoric correlation (.36) between the two predicted time series.

3.3 Bigram Distance in *To Build a Fire*: Proof of Concept

Table 5 reflects semantic distances for sequential and random bigrams in *To Build a Fire*. There was a very large effect of text structure (i.e., ordered wording) on embedding-based semantic distances. That is, sequential bigrams in the original text had significantly lower semantic distance relative to random bigram pairings [Welch-Satterthwaite $t(6333.2) = 57.61, p < .001$, Cohen's $d = 1.41$ (very large effect)]. This effect of text structure was weaker for experiential semantic distances (SemDist15) as evident by a small to medium effect for sequential relative to random bigrams [Welch-Satterthwaite $t(6381.2) = 12.57, p < .001$, Cohen's $d = .32$ (small to medium effect)].

Figure 3 illustrates the distribution of semantic distances across all bigrams of the original story. Note, the random bigram model reflected pairing of a target word in its original position within *To Build a Fire* with another word randomly selected from the SemDist15 database. The guidelines in Figure 2 reflect boundaries for low, medium, and high distance generated in the norming study (see Tables 2 & 3).

3.4 Semantic Distance for Bigrams across Sentence Boundaries in *To Build a Fire*

Semantic distances differed for sequential bigrams within sentences ($N = 5200$) relative to bigrams split across sentence boundaries ($N = 768$) across both the embedding and experiential models. Experiential (SemDist15) distance had a mean cosine distance normalized from 0–2 (0 is identical) for within-sentence bigrams (mean Cosine distance = .86) relative to the between-sentence (switch) condition (mean Cosine distance = .80), $t(487.89) = 3.09, p = .002$, Cohen $d = .18$, small effect.⁴ Embedding (GloWCA) distances were also slightly higher for between-sentence bigrams (mean cosine distance = .85) relative to within-sentence bigrams (mean distance = .82), $t(451.48) = 2.7, p = .006$, Cohen $d = .18$, small effect).

We conducted a replication analysis to determine whether the surprising finding of lower between-sentence than within-sentence semantic distance observed in *To Build a Fire* would hold across a much larger and more varied corpus.⁵ Using the same procedures applied to *To Build a Fire*, we analyzed bigram distances across ten novels (see Table 6) sourced primarily from Project Gutenberg and the Harry Potter R package (Boehmke. 2022). These texts included a total of approximately 906,421 words and 56,675 sentences. For 9 of the 10 sources, embedding-based semantic distance was *higher* for bigrams crossing sentence boundaries relative to within-sentence bigrams. For the remaining text (i.e., *The Portrait of Dorian Gray* by Oscar Wilde), distance was lower for between sentence bigrams, although this statistically significant difference constituted a very small effect. These overall results suggest that in most instances readers can expect to experience small jumps in semantic

⁴Distances scaled from the original range of –1:1 to 0:2 and reverse scored such that 0 constitutes identical vectors and 2 is the greatest possible distance between any pair of words.

⁵This suggestion was raised by an anonymous reviewer, highlighting the importance of replication and extension with largescale language models.

distance across sentence boundaries, as confirmed by a paired t -test on total between vs. within bigram distances (embedding-based)⁶ across all ten texts, $t(9) = 3.97$, $p = .002$. We interpret these results in the general discussion to follow.

4. General Discussion

Much remains to be learned about the ways that meaning is conveyed in continuous language, and recent advances in natural language processing have afforded new insights into the ways that words combine at different scales. Here we evaluated a sequential bigram model involving the application of two high dimensional semantic spaces to any continuous language sample. We developed open-source software for computing sequential bigram distance in a specified language sample of any length.⁷ We also derived semantic distance norms that serve as a standard against which other analyses of connected language might be gauged. In Section 3.2 we conducted a validation study demonstrating how both semantic spaces described in this work could successfully segment a continuous sample of simulated verbal fluency data. Finally, in sections 3.3 and 3.4, we evaluated word-to-word semantic transitions in *To Build a Fire*. We discuss some of the major findings to follow. Semantic distance is only interpretable in relation to the unique semantic space used to define it. For example, semantic distances between experiential and embedding models are not directly comparable. In addition, cosine values are not typically normalized, reverse scored, or standardized across semantic spaces. As such, when one observes a cosine similarity value of 0.6, it is almost impossible to determine the magnitude of this semantic distance in the absence of a known standard. In section 2.5, we described steps for establishing bigram distance norms for experiential (SemDist15) and embedding (GloWCA) spaces. We queried hundreds of thousands of naturally occurring bigrams in contemporary English text relative to ‘synthetic’ bigrams created by random word pairings. These analyses established bounds for low, medium, and high semantic distance which we then deployed as reference points for the analyses of *To Build a Fire*.

We used the norms established in section 2.5 to evaluate a proof-of-concept that bigram distance in *To Build a Fire* would be higher for random bigram pairings relative to naturally ordered text, a pattern that was also evident in the bigram norming sample. As predicted, adjacent words within *To Build a Fire* are more semantically related than a randomly sampled bigrams within the same corpus. Relatedness between words in running text was far stronger for embedding (GloWCA) than experiential (SemDist15) distances, suggesting dominance for thematic (e.g., dog-collar) relative to taxonomic (e.g., dog-wolf) semantic relatedness in discourse.

The nature of how thematic and taxonomic semantic systems interact during language comprehension remains one of the most active topics in cognitive science. It has been argued that embedding (i.e., thematic) models are also capable of recovering taxonomic relationships (Grand, Blank, Pereira, & Fedorenko 2022). However, the reverse pattern

⁶There were no reliable or systematic differences in bigram distance between and within sentences via the experiential semantic space (SemDist15)

⁷All associated code is freely available for inspection and use within the ‘semdistflow’ R-package. We encourage researchers to contact us for assistance.

appears less likely (i.e., experiential models recovering thematic relationships) since experiential models are more sensitive to perceptual than contextual similarity. Language discourse tends to unfold in terms of thematically related content (i.e., topics) making it more likely to encounter the word ‘leash’ than ‘wolf’ in proximity of ‘dog.’ This is not to say that taxonomic relationships are not present or important, particularly for scientific genres such as biology or zoology. However, our data suggest a more integral role for thematic relationships in online language processing.

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, excessively high semantic distance (i.e., each word highly unrelated to the last word) could be perceived as analogous to ‘word salad’ in terms of cohesion. In our first patient-based extension of this method, Litovsky et al. (2022) sampled hundreds of thousands of bigrams from the narratives of people with aphasia relative to age-matched controls. This bag-of-bigrams approach demonstrated that people with aphasia show reduced bigram semantic distances relative to controls and that compression in semantic distance strongly correlates with semantic ability. In the current project, we extended this approach to model semantic distance across ordered discourse, potentially expanding the power and ecological validity of the measure to treat conceptual drift across words as a time series.

We demonstrated several applications of this continuous bigram approach. First, we conducted a validation study examining whether two metrics of semantic distance could effectively mark cluster boundaries in a simulated stream of verbal fluency data. Both semantic spaces (GloWCA and SemDist15) showed higher than 90% classification accuracy, demonstrating sensitivity to detect semantic fluctuations in continuous language output (see also Zemla et al, 2020). Second, we tested whether bigrams spanning sentence boundaries had higher semantic distances than bigrams within sentences. We predicted that between-sentence bigrams would be marked by a jump in semantic distance relative to within-sentence bigrams. This is premised on the idea that sentences constitute micro-topics organized around thematic semantic content and that transitions between sentences incur associated shifts in meaning.

We initially found that semantic distance was paradoxically lower for bigrams crossing sentence boundaries in *To Build a Fire*. However, a subsequent replication/extension analysis revealed that this finding was likely idiosyncratic. A more extensive corpus analysis across several works of fiction revealed that the final content word of one sentence and the initial content word of the next sentence tend to be punctuated by a jump in semantic distance, consistent with our original hypothesis. The extent to which readers and listeners are sensitive to such jumps to facilitate online sentence parsing remains unclear.

4.2 Applications and Future Directions

We envision a variety of applications to conceptualizing language as a continuous time series fluctuating in meaning over time. This includes the following:

1. Causal Modeling of Language and Physiological Relations: 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 embedding semantic distance in a running narrative could tax cognitive control which in turn perturbs pupil diameter and heart rate. One fruitful extension of this method will involve continuous measurement of neural signals using techniques such as MEG or EEG with adequate temporal resolution to detect processes underlying predictive coding and the interplay between taxonomic and semantic systems during real-time language comprehension (for recent related work see Brodbeck, Hong, & Simon 2019; Kuperberg et al., 2006; Laszlo & Federmeier, 2009).

2. Implicit Detection of Neurological Disorders: 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 (Fraser et al., 2015; Gerard et al., 2001; Gerard et al., 2014). 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 algorithm that considers aberrant semantic distance in connected language as a marker of cognitive impairment.
3. Norming of Developmental Milestones: 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. Auto-segmenting Verbal Fluency Data: Our validation study demonstrates the utility of semantic distance in detecting switches between semantic clusters without the necessity for supervised machine learning or human intervention (e.g., manual scoring). This algorithm may prove useful for neuropsychology and other disciplines such as speech-language pathology that rely on verbal fluency as part of their core clinical assessment protocols.
5. Evaluating how semantic distance is moderated by part-of-speech: One of the key steps needed to refine the proposed model is to improve sensitivity to disambiguate grammatical class (e.g., run as a verb vs. run as a noun). Little is currently known about how bigram semantic distance is moderated by grammatical role and how semantic distance might contribute to thematic role assignment and verb argument structure.

4.3 Limitations

Language is a rich symbolic modality comprised of numerous interactive subdomains. In its current form, however, our algorithm can only yield coarse data about how the meaning of one word relates to its neighbor. Our processing pipeline is currently insensitive to grammatical, pragmatic, and/or lexical ambiguity. The program yokes each word to its single

entry in one of two lookup databases. The algorithm is agnostic to part of speech, polysemy, or homophony. This shortcoming undoubtedly results in error variance that could potentially be ameliorated by part-of-speech tagging or syntactic parsing. One rate limiting factor for the widespread adoption of such techniques involves the extensive processing resources required to parse large language samples. Such analyses often exhaust the capacity of personal computers, requiring database integration over high performance clusters. As such, refinement of the semantic algorithms proposed here will require integration of syntactic and pragmatic information to provide a more realistic picture of combinatorial semantic processing.

4.4 Conclusions

We have proposed a continuous bigram model and opensource toolkit for analyzing semantic transitions in natural language and have identified numerous applications of the model to address theoretical and clinical questions about combinatorial semantic processing. Much remains to be learned about how to best measure conceptual shifts in language and how such variability either facilitates or compromises human communication. We invite researchers to explore these tools with their own datasets.

Acknowledgement

We are grateful to Bonnie Zuckerman for her assistance with programming. We thank the creators of the Lancaster Sensorimotor Norms, AffectVec, and GloVe for developing open-source tools in the best spirit of science. We also thank Professor Mark Davies for his generosity in allowing us to distribute word embeddings trained on the CoCA. This work was supported by the US National Institute on Deafness and Other Communication Disorders [NIH/NIDCD, DC013063]. High performance computing resources supported by US National Science Foundation [162506] and the US Army Research Laboratory [W911NF-16-2-0189]. This work was also partially supported by the United States - Israel Binational Science Foundation [2021040].

Conflict of Interest and Funding Statement

The authors have no known conflict of interest. All tools developed here are open source. This work was funded by a grant (R01 DC013063) from the National Institute on Deafness and Other Communication Disorders (NIH/NIDCD)

Data Availability

All data and associated code are available via the Open Science Foundation

REFERENCES

- Alacam Ö, Schüz S, Wegrzyn M, Kißler J, & Zarriß S (2022). Exploring semantic spaces for detecting Clustering and Switching in Verbal Fluency. *Proceedings of the 29th International Conference on Computational Linguistics*, 178–191. <https://aclanthology.org/2022.coling-1.16>
- Allport DA (1985). Distributed memory, modular subsystems and dysphasia. In Newman SK & Epstein R (Eds.), *Current perspectives in dysphasia* (pp. 207–244). Churchill Livingstone.
- Anderson AJ, Binder JR, Fernandino L, Humphries CJ, Conant LL, Raizada RDS, Lin F, & Lalor EC (2019). An Integrated neural decoder of linguistic and experiential meaning. *Journal of Neuroscience*, 39(45), 8969–8987. 10.1523/JNEUROSCI.2575-18.2019 [PubMed: 31570538]
- Baldassano C, Chen J, Zadbood A, Pillow JW, Hasson U, & Norman KA (2017). Discovering event structure in continuous narrative perception and memory. *Neuron*, 95(3), 709–721. <https://doi.org/10/gbrrws> [PubMed: 28772125]
- Barsalou LW (1983). Ad hoc categories. *Memory & Cognition*, 11(3), 211–227. [PubMed: 6621337]

- Barsalou LW (1991). Deriving categories to achieve goals. *Psychology of Learning and Motivation*, 27, 1–64.
- Beaty RE, & Johnson DR (2021). Automating creativity assessment with SemDis: An open platform for computing semantic distance. *Behavior Research Methods*, 53(2), 757–780. [PubMed: 32869137]
- Beaty RE, Zeitlen DC, Baker BS, & Kenett YN (2021). Forward flow and creative thought: Assessing associative cognition and its role in divergent thinking. *Thinking Skills and Creativity*, 41, 100859.
- Beauchamp MS, Lee KE, Haxby JV, & Martin A (2002). Parallel visual motion processing streams for manipulable objects and human movements. *Neuron*, 34(1), 149–159. [PubMed: 11931749]
- Benoit K, Watanabe K, Wang H, Nulty P, Obeng A, Müller S, & Matsuo A (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30), 774. 10.21105/joss.00774
- Binder JR, Conant LL, Humphries CJ, Fernandino L, Simons SB, Aguilar M, & Desai RH (2016). Toward a brain-based componential semantic representation. *Cognitive Neuropsychology*, 1–45. 10.1080/02643294.2016.1147426 [PubMed: 27484776]
- Binney RJ, Ashaie SA, Zuckerman BM, 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. [PubMed: 29655024]
- Binney RJ, Zuckerman BM, Waller HN, Hung J, Ashaie SA, & Reilly J (2018). Cathodal tDCS of the bilateral anterior temporal lobes facilitates semantically-driven verbal fluency. *Neuropsychologia*, 111, 62–71. 10.1016/j.neuropsychologia.2018.01.009 [PubMed: 29337133]
- Bonner MF, & Epstein RA (2021). Object representations in the human brain reflect the co-occurrence statistics of vision and language. *Nature Communications*, 12(1), 4081. 10.1038/s41467-021-24368-2
- Brachman RJ (1983). What IS-A is and isn't: An analysis of taxonomic links in semantic networks. *Computer*, 16(10), 30–36.
- Brodbeck C, Hong LE, & Simon JZ (2018). Rapid transformation from auditory to linguistic representations of continuous speech. *Current Biology*, 28(24), 3976–3983. [PubMed: 30503620]
- Brysaert M, & New B (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4), 977–990. <https://doi.org/41/4/977> [PubMed: 19897807]
- Capitani E, Laiacona M, Mahon BZ, & 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. 10.1080/02643290244000266 [PubMed: 20957571]
- Christensen AP, & Kenett YN (2021). Semantic network analysis (SemNA): A tutorial on preprocessing, constructing, and analyzing semantic networks. *Psychological Methods*. 10.1037/met0000463
- Chrysikou EG (2006). When shoes become hammers: Goal-derived categorization training enhances problem-solving performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(4), 935. [PubMed: 16822159]
- Cree GS, & 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. [PubMed: 12825636]
- Crutch SJ, Troche J, Reilly J, & Ridgway GR (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. 10.3389/fnhum.2013.00186 [PubMed: 23720617]
- 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. [PubMed: 7198562]
- 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 WA, Huth AG, Griffiths TL, Gallant JL, & Theunissen FE (2017). The hierarchical cortical organization of human speech processing. *Journal of Neuroscience*, 37(27), 6539–6557. <https://doi.org/10/gbmfvcv> [PubMed: 28588065]
- Deniz F, Nunez-Elizalde AO, Huth AG, & Gallant JL (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. [PubMed: 31427396]
- 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 JT, Gonnerman LM, Andersen ES, & Seidenberg MS (1998). Category-specific semantic deficits in focal and widespread brain damage: A computational account. *Journal of Cognitive Neuroscience*, 10(1), 77–94. [PubMed: 9526084]
- Dilkina K, & Lambon Ralph MA (2012). Conceptual Structure within and between Modalities. *Frontiers in Human Neuroscience*, 6, 333. 10.3389/fnhum.2012.00333 [PubMed: 23293593]
- Farah MJ, & McClelland JL (1991). A computational model of semantic memory impairment: Modality specificity and emergent category specificity. *Journal of Experimental Psychology: General*, 120(4), 339–357. [PubMed: 1837294]
- Fernandino L, Binder JR, Desai RH, Pendl SL, Humphries CJ, Gross WL, Conant LL, & Seidenberg MS (2016). Concept representation reflects multimodal abstraction: A Framework for embodied semantics. *Cerebral Cortex*, 26(5), 2018–2034. 10.1093/cercor/bhv020 [PubMed: 25750259]
- Fernandino L, Tong J-Q, Conant LL, Humphries CJ, & Binder JR (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. 10.1073/pnas.2108091119 [PubMed: 35115397]
- Fraser KC, Lundholm Fors K, Eckerström M, Öhman F, & Kokkinakis D (2019). Predicting MCI status from multimodal language data using cascaded classifiers. *Frontiers in Aging Neuroscience*, 11, 205. 10.3389/fnagi.2019.00205 [PubMed: 31427959]
- Fraser KC, Meltzer JA, & Rudzicz F (2015). Linguistic Features Identify Alzheimer’s Disease in Narrative Speech. *Journal of Alzheimer’s Disease*, 49(2), 407–422. <https://doi.org/10/f72r27>
- Fu Z, Wang X, Wang X, Yang H, Wang J, Wei T, Liao X, Liu Z, Chen H, & Bi Y (2022). Different computational relations in language are captured by distinct brain systems. *Cerebral Cortex*, bhac117. 10.1093/cercor/bhac117
- 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 MA, Hodges JR, & Patterson K (2001). Prototypicality, distinctiveness, and intercorrelation: Analyses of the semantic attributes of living and nonliving concepts. *Cognitive Neuropsychology*, 18(2), 125–174. [PubMed: 20945209]
- Garrard P, Rentoumi V, Gesierich B, Miller B, & Gorno-Tempini ML (2014). Machine learning approaches to diagnosis and laterality effects in semantic dementia discourse. *Cortex*, 55, 122–129. 10.1016/j.cortex.2013.05.008 [PubMed: 23876449]
- 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 LM, Andersen ES, Devlin JT, Kempler D, & Seidenberg MS (1997). Double dissociation of semantic categories in Alzheimer’s disease. *Brain and Language*, 57(2), 254–279. [PubMed: 9126416]
- Grand G, Blank IA, Pereira F, & Fedorenko E (2022). Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature Human Behaviour*, 1–13. 10.1038/s41562-022-01316-8
- Gray K, Anderson S, Chen E, Kelly JM, Christian M, Patrick J, Huang L, Kenett YN, & Lewis K (2019). Forward flow: A new measure to quantify free thought and predict creativity. *American Psychologist*, 74(5), 539–554. 10.1037/amp0000391 [PubMed: 30667235]
- 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(3), 628–649. 10.1093/brain/awh075 [PubMed: 14761903]

- Grossman M, Peelle JE, Smith EE, McMillan CT, Cook P, Powers J, Dreyfuss M, Bonner MF, 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. 10.1016/j.neuroimage.2012.11.057 [PubMed: 23220494]
- 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. [PubMed: 31505121]
- Hartung F, Kenett YN, Cardillo ER, Humphries S, Klooster N, & Chatterjee A (2020). Context matters: Novel metaphors in supportive and non-supportive contexts. *NeuroImage*, 212, 116645. [PubMed: 32070752]
- Hillis AE, & Caramazza A (1991). Category-specific naming and comprehension impairment: A double dissociation. *Brain*, 114, Pt 5, 2081–2094. [PubMed: 1933235]
- Hills TT, & Kenett YN (2022). Networks of the mind: How can network science elucidate our understanding of cognition? *Topics in Cognitive Science*, 14(1):189–208. 10.1111/tops.12570. [PubMed: 34435461]
- Hills TT, Todd PM, & Jones MN (2015). Foraging in semantic fields: How we search through memory. *Topics in Cognitive Science*, 7, 513–534. 10.1111/tops.12151 [PubMed: 26097107]
- Hodges JR, 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](https://doi.org/S0093-934X(96)90077-2) [PubMed: 8811960]
- Hoffman P, McClelland JL, & Lambon Ralph MA (2018). Concepts, control, and context: A connectionist account of normal and disordered semantic cognition. *Psychological Review*, 125(3), 293. <https://doi.org/10/gdjrq9> [PubMed: 29733663]
- Hoffman P, Rogers TT, & Lambon Ralph MA (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. [PubMed: 21254804]
- Huth AG, De Heer WA, Griffiths TL, Theunissen FE, & Gallant JL (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600), 453–458. <https://doi.org/10/bfqp> [PubMed: 27121839]
- Huth AG, Nishimoto S, Vu AT, & Gallant JL (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6), 1210–1224. [PubMed: 23259955]
- Jackendoff R (1987). On Beyond Zebra: The relation of linguistic and visual information. *Cognition*, 26(2), 89–114. 10.1016/0010-0277(87)90026-6 [PubMed: 3652651]
- Jain S, & Huth AG (2018). Incorporating context into language encoding models for fMRI. *BioRxiv*, 327601.
- Johnson DR, Kaufman JC, Baker BS, Patterson JD, Barbot B, Green AE, van Hell J, Kennedy E, Sullivan GF, Taylor CL, Ward T, & Beaty RE (2022). Divergent semantic integration (DSI): Extracting creativity from narratives with distributional semantic modeling. *Behavior Research Methods*. 10.3758/s13428-022-01986-2
- Kacmador M, & Kelleher JD (2020). Capturing and measuring thematic relatedness. *Language Resources and Evaluation*, 54(3), 645–682. <https://doi.org/10/gn655n>
- Kalénine S, Mirman D, & Buxbaum LJ (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. 10.3389/fnhum.2012.00106 [PubMed: 22586383]
- Kalénine S, Peyrin C, Pichat C, Segebarth C, Bonthoux F, & Baciou M (2009). The sensory-motor specificity of taxonomic and thematic conceptual relations: A behavioral and fMRI study. *NeuroImage*, 44(3), 1152–1162. 10.1016/j.neuroimage.2008.09.043 [PubMed: 18977304]
- Kenett YN (2018). Investigating creativity from a semantic network perspective. In Kapoula Z, Volle E, Renoult J, & Andreatta M (Eds.), *Exploring Transdisciplinarity in Art and Science* (pp. 49–75). Cham: Springer
- Kenett YN (2019). What can quantitative measures of semantic distance tell us about creativity? *Current Opinion in Behavioral Sciences*, 27, 11–16. DOI: 10.1016/j.cobeha.2018.08.010

- Kenett YN, & Faust M (2019). Clinical cognitive networks: A graph theory approach. In Vitevitch MS (Ed.), *Network Science in Cognitive Science* (pp. 136–165). Routledge
- Kenett YN, Levi E, Anaki D, & Faust M (2017). The semantic distance task: Quantifying semantic distance with semantic network path length. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 43(9), 1470–1489. DOI: 10.1037/xlm0000391 [PubMed: 28240936]
- Kousta S-TT, Vigliocco G, Vinson DP, 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] [PubMed: 21171803]
- Kumar AA (2021). Semantic memory: A review of methods, models, and current challenges. *Psychonomic Bulletin & Review*, 28(1), 40–80. [PubMed: 32885404]
- Kumar AA, Balota DA, & Steyvers M (2020). Distant connectivity and multiple-step priming in large-scale semantic networks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(12), 2261–2276. 10.1037/xlm0000793 [PubMed: 31789562]
- Kumar AA, Steyvers M, & Balota DA (2022). A Critical review of network-based and distributional approaches to semantic memory structure and processes. *Topics in Cognitive Science*, 14(1), 54–77. 10.1111/tops.12548 [PubMed: 34092042]
- Kuperberg GR, Caplan D, Sitnikova T, Eddy M, & Holcomb PJ (2006). Neural correlates of processing syntactic, semantic, and thematic relationships in sentences. *Language and Cognitive Processes*, 21(5), 489–530. 10.1080/01690960500094279
- Landauer TK, & Dumais ST (1997). Solution to Plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211–240.
- Laszlo S, & Federmeier KD (2009). A beautiful day in the neighborhood: An event-related potential study of lexical relationships and prediction in context. *Journal of Memory and Language*, 61(3), 326–338. 10.1016/j.jml.2009.06.004 [PubMed: 20161064]
- 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 SJ (1979). The semantic nature of response competition in the picture-word interference task. *Memory & Cognition*, 7(6), 485–495.
- Lundin NB, Brown JW, Johns BT, Jones MN, Purcell JR, Hetrick WP, O’Donnell BF, & Todd PM (2022). Neural switch processes guide semantic and phonetic foraging in human memory. *PsyArXiv*. 10.31234/osf.io/857he
- Lynott D, Connell L, Brysbaert M, Brand J, & Carney J (2020). The Lancaster Sensorimotor Norms: Multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods*, 52(3), 1271–1291. 10.3758/s13428-019-01316-z [PubMed: 31832879]
- Maki WS, & 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> [PubMed: 18567261]
- Malt BC (2020). Words, thoughts, and brains. *Cognitive Neuropsychology*, 37(5–6), 241–253. [PubMed: 30964738]
- Malt BC, 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 BC, & Majid A (2013). How thought is mapped into words. *Wiley Interdisciplinary Reviews: Cognitive Science*, 4(6), 583–597. 10.1002/wcs.1251 [PubMed: 26304265]
- 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.
- Marelli M, Gagné CL, & Spalding TL (2017). Compounding as abstract operation in Semantic Space: A data-driven, large-scale model for relational effects in the processing of novel compounds. *Cognition*, 166, 207–224. [PubMed: 28582684]

- 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. 10.3758/s13423-015-0842-3 [PubMed: 24847901]
- 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](https://doi.org/S0010-0277(10)00254-4) [PubMed: 21093853]
- McRae K, Cree GS, Westmacott R, & Sa VRD (1999). Further evidence for feature correlations in semantic memory. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, 53(4), 360. [PubMed: 10646207]
- Merkin A, Krishnamurthi R, & Medvedev ON (2022). Machine learning, artificial intelligence and the prediction of dementia. *Current Opinion in Psychiatry*, 35(2), 123–129. 10.1097/YCO.0000000000000768 [PubMed: 34861656]
- Mikolov T, Sutskever I, Chen K, Corrado GS, & 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 AE (2017). Taxonomic and thematic semantic systems. *Psychological Bulletin*, 143(5), 499. <https://doi.org/10/gkbqjp> [PubMed: 28333494]
- Mota NB, 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. 10.1038/s41537-017-0019-3 [PubMed: 28560247]
- Mota NB, Vasconcelos NA, Lemos N, Pieretti AC, Kinouchi O, Cecchi GA, Copelli M, & Ribeiro S (2012). Speech graphs provide a quantitative measure of thought disorder in psychosis. *PLoS One*, 7(4), e34928. 10.1371/journal.pone.0034928 [PubMed: 22506057]
- Naselaris T, Kay KN, Nishimoto S, & Gallant JL (2011). Encoding and decoding in fMRI. *NeuroImage*, 56(2), 400–410. [PubMed: 20691790]
- Nastase SA, Goldstein A, & Hasson U (2020). Keep it real: Rethinking the primacy of experimental control in cognitive neuroscience. *NeuroImage*, 222, 117254. [PubMed: 32800992]
- Olson JA, Nahas J, Chmoulevitch D, Cropper SJ, & Webb ME (2021). Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences*, 118(25), e2022340118. 10.1073/pnas.2022340118
- Ovando-Tellez M, Benedek M, Kenett YN, Hills TT, Bernard M, Belo Y, Belanger B, Bieth T, & Volle E (2022). An investigation of the cognitive and neural correlates of semantic memory search related to creative ability. *Communications Biology*, 5(1), 604. 10.1038/s42003-022-03547-x [PubMed: 35710948]
- 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> [PubMed: 17400539]
- Pennington J, Socher R, & Manning CD (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 ME, Ammar W, Bhagavatula C, & Power R (2017). Semi-supervised sequence tagging with bidirectional language models. *ArXiv Preprint ArXiv:1705.00108*.
- Pexman PM, Heard A, Lloyd E, & Yap MJ (2016). The Calgary semantic decision project: Concrete/abstract decision data for 10,000 English words. *Behavior Research Methods*, 1–11. 10.3758/s13428-016-0720-6 [PubMed: 25761390]
- Popham SF, Huth AG, Bilenko NY, Deniz F, Gao JS, Nunez-Elizalde AO, & Gallant JL (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> [PubMed: 34711960]
- Prabhakaran R, Green AE, & Gray JR (2014). Thin slices of creativity: Using single-word utterances to assess creative cognition. *Behavior Research Methods*, 46(3), 641–659. 10.3758/s13428-013-0401-7 [PubMed: 24163211]
- Price AR, Bonner MF, Peelle JE, & Grossman M (2015). Converging evidence for the neuroanatomic basis of combinatorial semantics in the angular gyrus. *The Journal of Neuroscience*, 35(7), 3276–3284. 10.1523/JNEUROSCI.3446-14.2015 [PubMed: 25698762]

- Raji S, & da Melo G (2020). What sparks joy: The AffectVec emotion database. Proceedings of the Web Conference, ACM.
- Reilly J, Finley AM, 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 JE, Garcia A, & Crutch SJ (2016). Linking somatic and symbolic representation in semantic memory: The dynamic multilevel reactivation framework. *Psychonomic Bulletin & Review*, 23(4), 1002–1014. 10.3758/s13423-015-0824-5 [PubMed: 27294419]
- Reilly J, Kelly A, Zuckerman BM, Twigg PP, Wells M, Jobson KR, & Flurie M (2020). Building the perfect curse word: A psycholinguistic investigation of the form and meaning of taboo words. *Psychonomic Bulletin & Review*. 10.3758/s13423-019-01685-8
- Reilly J, Zuckerman B, Litovsky C, & Finley A (2021). semdistflow: Strip, format, vectorize text to sequential bigram semantic distance (1.0) [Computer software]
- Rinker TW (2018a). Textstem: Tools for stemming and lemmatizing (0.1.4) [Computer software].
- Rinker TW (2018b). Textclean (0.9.3) [Computer software]. <https://CRAN.R-project.org/package=textclean>
- Rogers TT, Lambon Ralph MA, Garrard P, Bozeat S, McClelland JL, Hodges JR, & Patterson K (2004). Structure and deterioration of semantic memory: A Neuropsychological and computational investigation. *Psychological Review*, 111(1), 205–235. [PubMed: 14756594]
- Rosch EH (1973). Natural categories. *Cognitive Psychology*, 4(3), 328–350. 10.1016/0010-0285(73)90017-0
- Sacchett C, & Humphreys GW (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 JL, & 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. 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. 10.1016/j.concog.2020.103070 [PubMed: 33307427]
- Schwartz MF, Kimberg DY, Walker GM, Brecher A, Faseyitan OK, Dell GS, Mirman D, & Coslett HB (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 CJ, 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. 10.1212/WNL.0b013e318225c6bc [PubMed: 21753161]
- Spooner A, Chen E, Sowmya A, Sachdev P, Kochan NA, 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. 10.1038/s41598-020-77220-w [PubMed: 31913322]
- Troche J, Crutch SJ, & 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. 10.3389/fpsyg.2017.01787
- Troyer AK, Moscovitch M, Winocur G, Alexander MP, & Stuss D (1998). Clustering and switching on verbal fluency: The effects of focal frontal- and temporal-lobe lesions. *Neuropsychologia*, 36(6), 499–504. [PubMed: 9705059]
- Troyer AK (2000). Normative data for clustering and switching on verbal fluency tasks. *Journal of Clinical and Experimental Neuropsychology*, 22(3), 370–378.

- Trumpp NM, 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. [PubMed: 22405961]
- Warriner AB, Kuperman V, & Brysbaert M (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4), 1191–1207. 10.3758/s13428-012-0314-x [PubMed: 23404613]
- Wang X, Men W, Gao J, Caramazza A, & Bi Y (2020). Two forms of knowledge representations in the human brain. *Neuron*, 107(2), 383–393. 10.1016/j.neuron.2020.04.010 [PubMed: 32386524]
- Warrington EK (1975). The selective impairment of semantic memory. *Quarterly Journal of Experimental Psychology*, 27(4), 635–657. [PubMed: 1197619]
- Westerlund M, & Pyllkänen L (2014). The role of the left anterior temporal lobe in semantic composition vs. Semantic memory. *Neuropsychologia*, 57, 59–70. 10.1016/j.neuropsychologia.2014.03.001 [PubMed: 24631260]
- Wingfield C, & Connell L (2022). Sensorimotor distance: A grounded measure of semantic similarity for 800 million concept pairs. *Behavior Research Methods*. 10.3758/s13428-022-01965-7
- Wollams AM, Cooper-Pye E, Hodges JR, & Patterson K (2008). Anomia: A doubly typical signature of semantic dementia. *Neuropsychologia*, 46(10), 2503–2514. [PubMed: 18499196]
- 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. 10.18637/jss.v007.i02
- Zemla JC, Cao K, Mueller KD, & Austerweil JL (2020). SNAFU: The semantic network and fluency utility. *Behavior Research Methods*, 52(4), 1681–1699. 10.3758/s13428-019-01343-w [PubMed: 32128696]
- 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. 10.1016/j.neuroimage.2020.117405 [PubMed: 32992002]

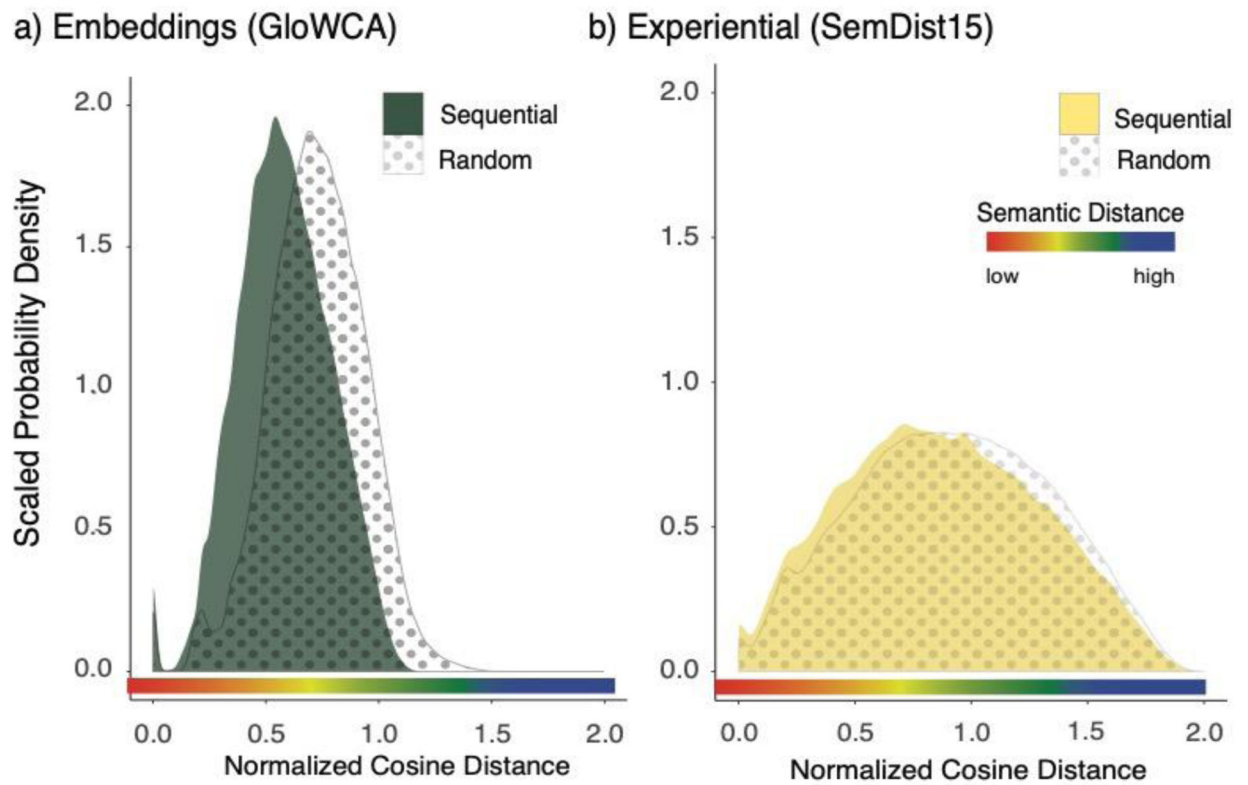


Figure 1. Semantic Distance Density Distributions for Sequential and Random Bigrams

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 .

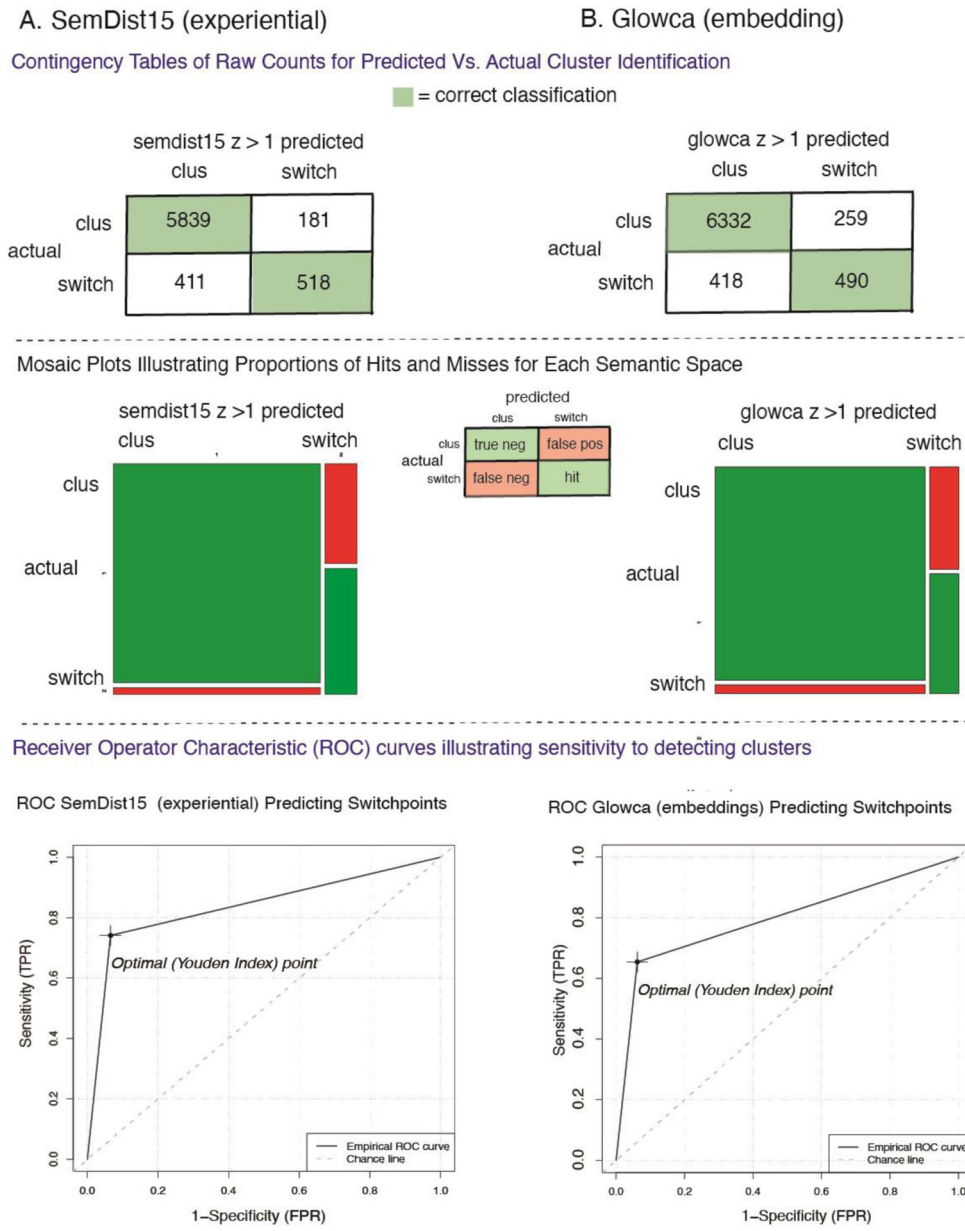
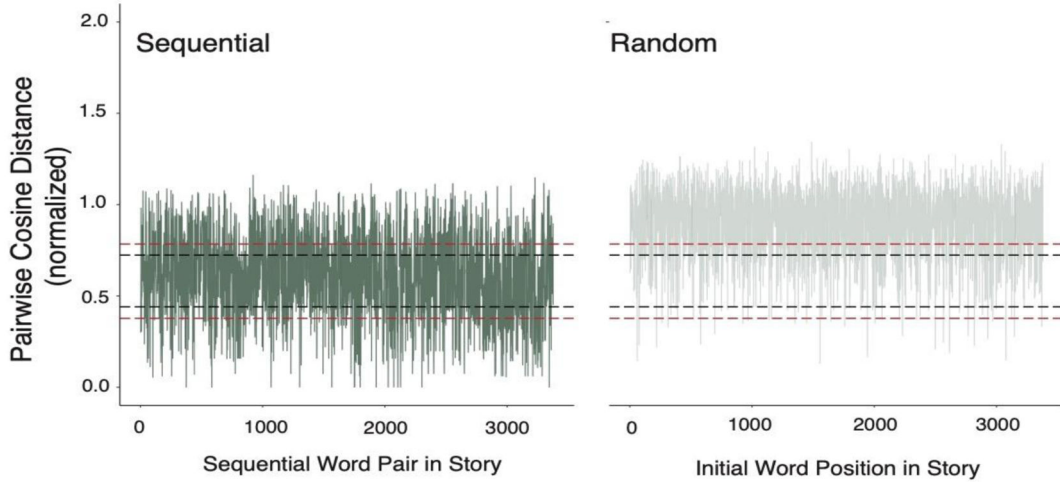


Figure 2. Sensitivity and specificity of binary classification of simulated category fluency

a) Embedding Distance (GloWCA)



b) Experiential Distance (SemDist15)

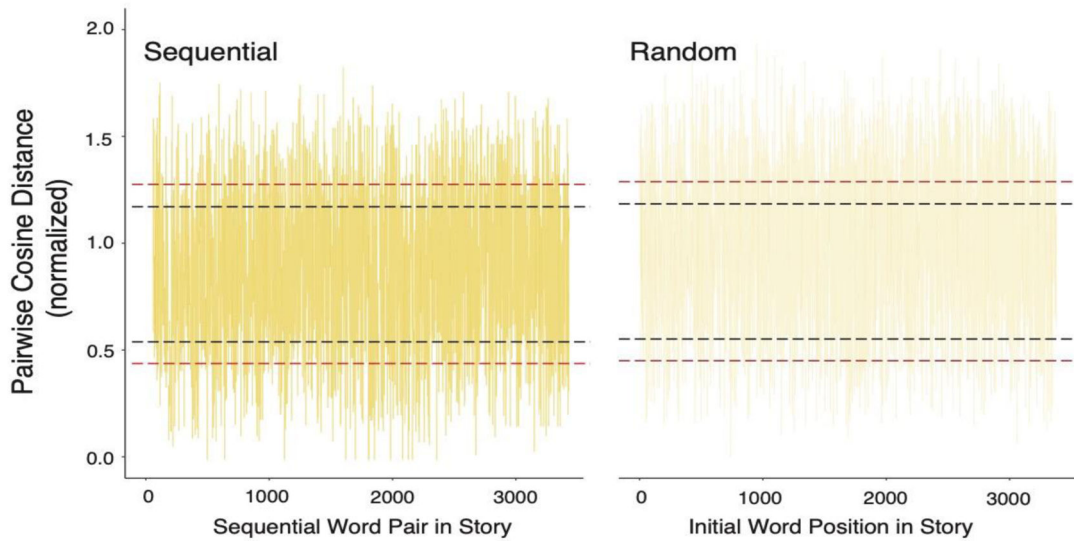


Figure 3. Semantic Distances in To Build a Fire

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.

Table 1.

Text Stripping, Vectorizing, and Global Formatting

Target	Description of Global Action (substitution or omission)
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).

Table 2.

SemDist15 Experiential Bigram Distances: Norms and Ranges

		Distance Band	Bigram Condition	
			Sequential	Random
Mean (SD)	Raw Cosine (-1:1)	n/a	0.13 (0.42)	0.08
	Rescaled Reverse Scored (0:2)		0.87 (0.42)	0.92 (0.42)
Min to Q1	Raw Cosine (-1:1)	Low	0.45 to 1.00	0.39 to 1.00
	Rescaled Reverse Scored (0:2)		0 to 0.55	0 to 0.61
Q1 to Q2	Raw Cosine (-1:1)		0.14 to 0.44	0.08 to 0.38
	Rescaled Reverse Scored (0:2)		0.56 to 0.86	0.62 to 0.92
Q2 to Q3	Raw Cosine (-1:1)	Average	-0.18 to 0.13	-0.76 to 0.07
	Rescaled Reverse Scored (0:2)		0.87 to 1.18	0.93 to 1.24
Q3 to Max	Raw Cosine (-1:1)	High	-0.96 to -0.19	-0.97 to 0.75
	Rescaled Reverse Scored (0:2)		1.19 to 1.96	1.25 to 1.97

Note: Raw Cosine scores reflect the original cosine value on a -1 to 1 scale with a cosine of 1 indicating 0 distance between two vectors. Rescaled Reverse Scored (0:2) values reflect a transformation of the original cosine values first to a range between zero to two and then reverse scored using the 1-observed. On this transformed scale, zero indicates no distance (i.e., a word vs. itself), and two reflects the highest possible dissimilarity between two words.

Table 3.

Glowca Embedding Bigram Distances: Norms and Ranges

		Distance Band	Bigram Condition	
			Sequential	Random
Mean (SD)	Raw Cosine (-1:1)	n/a	0.42	0.28
	Rescaled Reverse Scored (0:2)		0.58 (0.20)	0.72 (0.22)
Min to Q1	Raw Cosine (-1:1)	Low	0.56 to 1.00	0.41 to 1.00
	Rescaled Reverse Scored (0:2)		0 to 0.44	0 to 0.59
Q1 to Q2	Raw Cosine (-1:1)		0.42 to 0.55	0.27 to 0.40
	Rescaled Reverse Scored (0:2)		0.45 to 0.58	0.60 to 0.73
Q2 to Q3	Raw Cosine (-1:1)	Average	0.27 to 0.41	0.13 to 0.26
	Rescaled Reverse Scored (0:2)		0.59 to 0.73	0.74 to 0.87
Q3 to Max	Raw Cosine (-1:1)	High	-0.22 to 0.26	-0.59 to 0.12
	Rescaled Reverse Scored (0:2)		0.74 to 1.22	0.88 to 1.59

Note: Raw Cosine scores reflect the original cosine value on a -1 to 1 scale with a cosine of 1 indicating 0 distance between two vectors. Rescaled Reverse Scored (0:2) values reflect a transformation of the original cosine values first to a range between zero to two and then reverse scored using the 1-observed. On this transformed scale, zero indicates no distance (i.e., a word vs. itself), and two reflects the highest possible dissimilarity between two words.

Table 4.

Accuracy of binary classification for simulated category fluency

	SemDist15	Glowca
% Accuracy	91.5%	90.9%
d'	2.02	1.86
Hit Rate	93.4%	93.80%
False Alarm Rate	6.58%	6.20%
Odds Ratio	40.66	28.66
Bias	0.96	0.98

Note: Sensitivity metrics derived from the 'verification' and 'psych' packages in R.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 5.Semantic Distances for *To Build a Fire*

	Condition	N	Mean Dist	SD
Semdist15	Sequential	3134	0.90	.36
	Randomized	3255	1.01	.36
GloWCA	Sequential	3360	0.62	.23
	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.

Table 6.

Embedding distance for bigrams within sentences vs. crossing sentence boundaries

Source	Token Counts		Bigram Distance		t-statistic
	Words	Sentences	Within	Between	
Prisoner of Azkaban	104860	8936	.62	.65	t(7514.9) = 7.92, p < .001 ***
Little Women	194059	9266	.61	.63	t(8498) = 3.55, p < .001 ***
Sherlock Holmes	107372	7065	.60	.61	t(7335) = 3.01, p = .002 **
Portrait of Dorian Gray	82012	6687	.60	.59	t(7150) = 3.77, p < .001 ***
Pride and Prejudice	124719	6210	.61	.62	t(6569.6) = 2.33, p = .02*
Room with a View	69931	5948	.62	.63	t(6192.3) = 2.79, p = .01*
Sorcerer's Stone	77536	6474	.60	.63	t(5874.7) = 5.46, p < .001 ***
Become an Engineer	21072	1466	.64	.68	t(1694.6) = 6.92, p < .001 ***
Honey Bees	91577	3182	.72	.75	t(3061.3) = 8.97, p < .001 ***
Prehistoric Villages	33283	1541	.67	.72	t(1521) = 7.05, p < .001 ***

Note:

**
p < .01,***
p < .001;

Token counts derived using the Quanteda package of R (Benoit et al, 2018). Distances reflect 0–2 cosine rescaled and reverse scored (0 is identical). Texts queried: Harry Potter and the Prisoner of Azkaban (J.K. Rowling, 1999); Little Women (Louisa May Alcott, 1868); The Adventures of Sherlock Holmes (Arthur Conan Doyle, 1892); The Portrait of Dorian Gray (Oscar Wilde, 1890); Pride and Prejudice (Jane Austen, 1813); A Room with a View (E.M. Forster, 1908); Harry Potter and the Sorcerer's Stone (J.K. Rowling, 1998); How to Become an Engineer (Frank W. Doughty, 2014); The Honey Bee: Its Natural History, Physiology, and Management (Edward Bevens, 1873); Prehistoric Villages, Castles, and Towers of Southwestern Colorado (Jesse Fewkes, 1919)