Spatial vs. Graphical Representation of Distributional Semantic Knowledge

Shufan Mao, Philip A. Huebner, and Jon A. Willits

Department of Psychology, University of Illinois at Urbana Champaign

Author Note:

This work has been presented as a poster at the 42nd Annual Meeting of the Cognitive Science Society, and a prior version of this manuscript has been published online at https://doi.org/10.31234/osf.io/p5fyv. All code, relevant results and supplementary materials for the paper and the available at https://github.com/UIUCLearningLanguageLab/Humans.

We have no conflicts to disclose.

This study was not preregistered.

We appreciate Dr. Gary Dell and Dr. Kara Federmeier for suggestions on the article, and all members of the Learning and Language Lab at UIUC for help and advices.

Correspondence concerning this article should be directed to Shufan Mao, University of Illinois at Urbana Champaign, 603 East Daniel St., Champaign, IL 61820, United States
Email: smao9@illinois.edu

Abstract

Spatial distributional semantic models represent word meanings in a vector space. While able to model many basic semantic tasks, they are limited in many ways, such as their inability to represent multiple kinds of relations in a single semantic space, and to directly leverage indirect relations between two lexical representations. To address these limitations, we propose a distributional graphical model that encodes lexical distributional data in a graphical structure and uses spreading activation for determining the plausibility of word sequences. We compare our model to existing spatial and graphical models by systematically varying parameters that contributing to dimensions of theoretical interest in semantic modeling. In order to be certain about what the models should be able to learn, we trained each model on an artificial corpus describing events in an artificial world simulation containing experimentally controlled verb-noun selectional preferences. The task used for model evaluation requires recovering observed selectional preferences and inferring semantically plausible but never observed verb-noun pairs. We show that the distributional graphical model performed better than all other models. Further, we argue that the relative success of this model comes from its improved ability to access the different orders of spatial representations with the spreading activation on the graph, enabling the model to infer the plausibility of noun-verb pairs unobserved in the training data. The model integrates classical ideas of representing semantic knowledge in a graph with spreading activation, and more recent trends focused on extraction of lexical distributional data from large natural language corpora.

keywords: semantic models, semantic network, distributional models, language comprehension, graphical models

Spatial vs. Graphical Representation of Distributional Semantic Knowledge

Representing and processing semantic information is fundamental to language. The sequence '*babies sleep*' is more easily processed than '*cars sleep*', and the sequence '*ideas sleep*' is even more difficult to process. Because all three sequences are grammatical and share the same syntactic structure, it is most natural to explain the difference in ease of processing at the semantic level. While neither cars nor ideas can sleep, it is easier for most people to metaphorically imagine a car sleeping than an idea sleeping. This example illustrates that semantic relatedness between words is an important part of people's ability to process and understand language.

The large number of words in natural languages, and the number of different ways that words can be related and paired, presents a daunting challenge for modeling semantic relatedness. Distributional models of semantic memory have been quite successful at modeling coarse-grained semantic tasks such as categorization (Landauer, Foltz & Laham, 1998; Lund & Burgess, 1996; Huebner & Willits, 2018) and semantic priming (Griffiths, Steyvers & Tenenbaum, 2007; Hutchison, Balota, Cortese, & Watson, 2007; Kumar et al., 2020; Landauer, Foltz & Laham, 1998; Mandara, Keuleers, & Brysbaert, 2017). And despite the success and wide applicability of both co-occurrence-based vector space models and more recent neural network models like Word2Vec (Mikolov et al., 2013), BERT (Devlin et al., 2018) and GPT-3 (Brown et al. 2020), these models still have known shortcomings. For example, these models often fail at rudimentary language tasks involving structured relations or compositionality (Lake & Murphy, 2021; Gershman & Tenenbaum, 2015; Marcus, 2020). And while the semantic coherence of large language models like GPT-4 is truly impressive, the fact that they require orders of magnitude more data in order to achieve that performance raises serious questions about their feasibility as models of human semantic representation. As a consequence, questions remain about the capacity of these kinds of models to represent the various kinds of semantic relations that humans use to represent and comprehend language.

In this paper, we address the question of what kind of semantic representations and processes might best support the human ability to produce graded semantic plausibility judgments of multi-word sequences. We discuss the advantages and disadvantages of representing word co-occurrence information and word similarity information in high-dimensional vector spaces, versus other approaches that represent this information in a connected graph (Anderson & Bower, 1974; Collins & Loftus, 1975; De Deyne, Navarro, Perfors & Storms, 2016; Gentner, 1975; Rotaru, Vigliocco, & Frank, 2018; Rumelhart & Levin, 1975). We then present a set of experiments designed to test the capacities of these models at predicting quantitative differences in semantic plausibility of predicate-argument pairs.

The paper is organized as follows: First, we describe the importance of lexical semantic relatedness in determining the semantic plausibility of a sentence. Second, we review the distributional approach to semantic modeling, and describe some features or properties that differentiate how lexical semantic relatedness is acquired by different semantic models. The major properties we examined are the representational structure of the model (graphical vs. spatial), and the type of information that is encoded (co-occurrence vs. similarity). Third, we discuss the construction of an artificial corpus that we built for the purpose of training and evaluating our models. Fourth, we report model performances in a selectional preference task that requires learning of semantic relations instantiated in the training data and making inferences about unobserved relations. Fifth, we explore the individual contributions of representational structure and information encoding type and their interaction to performance in this task. Finally, we discuss our results more broadly in the context of models of semantic development.

Simple Sentences and Syntagmatic Relatedness

The study of semantic knowledge representation can be approached from multiple perspectives. One important starting point is establishing what behavioral or empirical phenomenon one is trying to model or explain. Historically, researchers have examined a diverse array of phenomena, including the learning of word meanings, semantic priming, categorization and typicality effects, judgements about factuality or plausibility, and sentence production and comprehension. In this paper, we focused on how representations and processes underlying lexical semantic relatedness contribute to plausibility judgments of multi-word sequences.

There have been many proposals for how to characterize the nature and kinds of relations that can affect the semantic plausibility of multi-word sequences. One of the most foundational distinctions is between words that have a *syntagmatic* relationship and words that have a *paradigmatic* relationship (Saussure, 1983; Sahlgren, 2006). Simply put, words that are syntagmatically related are words that can "go together" in language, operationally defined as linguistic co-occurrence or thematic relatedness. Syntagmatic relatedness is most often and most easily used to describe noun-verb relations (*drink-coffee*, *walk-dog*), but can also be used to describe adjective-noun relations (*hot-coffee*, *brown-dog*) and noun-noun relations (*cup-coffee*, *leash-dog*) where relatedness is defined as co-occurrence or joint participation in the same event. Syntagmatic relatedness can be distinguished from *paradigmatic* relatedness, which links words that are substitutable with one another in the linguistic structures in which they occur. In the previous examples, *tea* is paradigmatically related to *coffee*, because *tea* and *coffee* can be substituted with minimal impact on the meaning of the sentences or their semantic plausibility.

Syntagmatic and paradigmatic relations are related concepts. For instance, under many theories (like those using some form of distributional learning), syntagmatic relatedness can be inferred based on a combination of paradigmatic and syntagmatic relatedness. For instance, based on previous knowledge of the paradigmatic relatedness of *dog-puppy*, and the syntagmatic relatedness of *dog-leash*, one can infer that *puppy* and *leash* are also likely to be syntagmatically related.

When judging the semantic plausibility of sentences, syntagmatic relatedness is probably more influential. The sentences '*Mary drank the coffee*' and *'Mary walked the dog'* are plausible sentences, and '*Mary drank the dog*' and *'Mary walked the coffee'* are not, because they mismatch in their syntagmatic rather than paradigmatic relations. In other words, it seems to make more sense to describe the plausibility of sentences in terms of how well the words *drink* and *coffee* go together (compared to *drink* and *dog*), and not in terms of their substitutability.

Selectional Preference

In this work, we were interested in how well different distributional semantic models acquire the selectional constraints on noun-verb pairs from linguistic data. We will use the term 'selectional preference' to describe constraints that determine which word pairs result in semantically plausible combinations (e.g. *'babies sleep'*) and which do not (e.g. *'ideas sleep'*). For example, *baby* is a better argument than *idea* for the verb *sleep*. Selectional preference is a quantitative (i.e. graded) phenomenon, and therefore requires a numerical score for each predicate-argument pair (Erk, S. Padó & U. Padó, 2010). Thus, it is critical that we derive model judgments on a continuous as opposed to discrete ("related vs. unrelated") scale.

We can derive a measurement of selectional preference from most semantic models equipped with quantitative measures of lexical relatedness. For example, many models represent words as vectors, and their relatedness as distances in a vector space (Osgood, 1957; Deese, 1962; Smith, Shoben, & Rips, 1974; Griffiths, Steyever, & Tenenbaum, 2007; Landauer & Dumais, 2007; Jones & Mewhort, 2007; Mikolov et al., 2013, Huebner & Willits, 2018). Other models represent semantic relations in a graphical structure (network or tree-like), with connections that vary in strength, and/or with a spreading activation mechanism that allows for a quantitative degree of relatedness between words (Collins & Quillian, 1969, Collins & Loftus, 1975, Elman, 1990; McRae, de Sa & Seidenberg, 1997 Miller 1995, Nelson, McEvoy & Schreiber, 2004, Steyvers & Tenenbaum 2005; Rumelhart & Todd, 1993; Rogers & McClelland, 2004). These models, despite varying considerably in their operational definitions of semantic relatedness, all provide a way to measure the relatedness between arbitrary word pairs. Therefore, it is possible to use model-derived relatedness scores as a proxy for selectional preference.

Research Goals

In this work, we use model-derived selectional preference judgments to evaluate the representational capabilities of different distributional semantic models. Importantly, to perform well in our evaluation, a model must not only assign higher semantic relatedness to word pairs that frequently co-occur, but also to word pairs that are more semantically plausible despite not having been observed during training (e.g. '*cars sleep*' is more plausible than '*ideas sleep*'). By comparing a model's selectional preferences to the corpus on which it was trained, we can make inferences about which models or properties of models are most useful for acquiring syntagmatic knowledge, and deploying that knowledge to make inferences about semantic plausibility.

Broadly, our work is a systematic comparison of many distributional models of semantics, with a specific focus on their ability to represent syntagmatic relations and using their learned representations to infer the semantic plausibility of observed and novel word sequences. There are four major differences between our approach and that of previous studies. First, our work systematically explores differences between graphical models built from language-internal distributional data and more traditional spatial models built on the same amount and kind of data. While several graphical models have been proposed in the semantic modeling literature, they have rarely been compared to spatial models while systematically controlling for differences in their training data, learning algorithm, and other modeling parameters. Second, we conducted a systematic comparison of model properties and parameters to better understand their individual contributions and their interactions. The third difference is a focus on the quantitative rather than qualitative nature of semantic relatedness. Previous work has focused on qualitative analyses, such as distinguishing "related" versus "unrelated" word pairs (Erk, S. Padó & U. Padó, 2010; Huebner & Willits, 2018; Bullinaria & Levy, 2007; Bullinaria & Levy, 2012). However, in this work we are more interested in the ability of models to reproduce a gradient of selectional preference that can be used to rank-order multiple word pairs by semantic plausibility. For example, given the pairs *'trap rabbit'* (observed), *'trap boar'* (unobserved, more plausible), *'trap water'* (unobserved, less plausible), previous evaluations required that a model only produce relatedness scores that differentiate the observed pairs from the unobserved. That is, a model only needed to correctly judge *'trap rabbit' > 'trap boar'* and, separately, *'trap rabbit' > 'trap water'*. In contrast, to succeed in our evaluation, a model must produce the correct rank-ordering *'trap rabbit' > 'trap boar' > 'trap water'*.

The fourth, and perhaps biggest difference between the present work and previous work, is the employment of carefully controlled artificial corpora to explore the differences between the models. In constrast, previous work has focused on training models on naturalistic linguistic corpora. While this has obvious benefits (e.g. providing the semantic model with input that humans are likely to experience), it leaves open the question of how to evaluate whether a model has come up with the "right" evaluation of relatedness. If two models are trained on the same corpora, and one model says that the most syntagmatically-related verb for *dog* should be *barked* and another model says *chased*, which model is correct? This can be an especially difficult problem to solve when one considers that the training corpus, too, is an implicit part of the model. Without knowing if the training corpus is a representative sample of the learning environment, it is difficult to assess which model has made more accurate predictions about that environment. In the current work, we follow in the path of previous work that addresses this problem by utilizing an artificial language approach to understand model dynamics (Elman, 1991; Rohde, D., 2002; Rubin, Kievit-Kylar, Willits, & Jones, 2014; Mao, Huebner, & Willits, 2022).

The corpus we used to train our models was specifically constructed for the purpose of evaluating semantic plausibility: Each sentence describes an event in a simulated world of hunter-gatherers that evolves according to deterministic rules. As the simulated world progresses, sentences are to describe events in that world using English words and pseudo-English grammatical rules. Importantly, because we created the set of rules that progress the simulated world from which the corpus is created, we have full access to knowledge about which possible noun-verb pairs are semantically more plausible. Put differently, we know precisely what the ground truth is with regard to semantic relations between nouns and verbs, and we can use this knowledge to test which models are better at recovering this structure from the corpus.

Semantic Modeling Theoretical Parameter Space

When semantic models are used to predict data and their performance is compared, it is not always clear why a particular model performed better than another. Unlike well-controlled experiments, computational models are complex, and usually vary in many ways at once.

To assist us in thinking about this issue, it can be valuable to think of semantic models themselves as existing in a multi-dimensional theoretical property space (Bullinaria & Levy, 2007, 2012; Rubin et al., 2014). While some properties of models have received more attention than others, each can impact a model's representational capabilities, often in unpredictable ways. Also, lack of understanding about the contribution of each property of a model may lead to attributing a model's successes and failures to incorrect factors (Jones, Gruenenfelder & Recchia, 2011). To provide a sense of the vastness of the theoretical space, and deeper insight into parametric relationships between existing semantic models, we list several semantic models and their choices on six important theoretical dimensions. These are shown in Table 1. It is important to note that these are not the only ways in which these models can vary. But the six dimensions highlight ways that semantic models tend to vary in theoretically interesting ways that affect what the models can and cannot do well.

The first theoretical dimension is the **Information Source**. This is the nature of the input used to build the model. There have traditionally been three sources from which to derive the information used to build a semantic model: 1) hand-labeled relations between words and concepts, chosen by the model's creator either for theoretical or demonstrative reasons, 2) normative relations, where relations between words are derived from empirically obtained normative experiments, such as word association norms (Nelson, McEvoy & Schreiber, 2004) or semantic feature norms (McRae, Cree, Seidenberg & McNorgan, 2005), and 3) linguistic corpora, where relations are derived from linguistic data.

The second theoretical dimension is the model's **Representational Structure**. This is the kind of data structure in which the semantic information is encoded. There have been two kinds of structures used by the vast majority of semantic models: 1) graphical models like structured trees (Collins & Quillian, 1969; Xu & Tenenbaum, 2007) and spreading activation on networks (Anderson, 1983; Collins & Loftus, 1975), and 2) vector spaces where each word or concept is represented as a point in some high-dimensional space, such as Osgood's Semantic Differential (1957), Smith, Shoben, and Rip's Feature Model (1974), and distributional language models like Latent Semantic Analysis (Landauer & Dumais, 1997) and Word2Vec (Mikolov et al., 2013).

With regard to the graph vs. space distinction, one potentially confusing class of models is connectionist and neural network models. Neural networks are, by any reasonable definition, graphical models. They have nodes, edges, and an algorithm for how activation spreads across the nodes. However, a distinction we would like to draw regarding whether a model is a spatial or a graphical model is how it is used when determining the relatedness between two words. If this is accomplished by applying the model's algorithm for spreading activation among the nodes and using a node's activity level as the measure of relatedness between two words/concepts, then the model should be thought of as a graphical model. For example, in the network studied by Rogers and McClelland (2004), when the inputs '*canary'* and '*is'* were activated, this would activate the output node *yellow*. In this model, the flow of activation from input to output is conceptually similar to spreading activation models. Likewise, in a next-word prediction model like the simple recurrent network (SRN, Elman, 1990) or GPT models (Brown et al., 2020; Radford et al., 2019), when an input is activated, and the predicted output activation is used as the measure of relatedness between the input and output, the model is properly thought of as a graph. However, some would argue that some neural network models are better thought of as a space. For example, in the Word2Vec model (Mikolov, 2013), the learned weights of a word in the network are used as a vector representation of words, which enables the computation of similarity between two words in terms of the similarity of their vector representations. Similarly, Huebner and Willits (2018) used the hidden state activations learned by the SRN to compute relatedness between word pairs. In such cases, it is more appropriate to think of these models as vector space models. While the vectors may have been derived from an algorithmic process on a graphical model, the resulting representation, and how it is used to determine similarity, is the same as in other spatial models. Thus, whether a connectionist or neural network model is more properly thought of as a graph or as a space depends on how the model is being used to define a construct of interest (e.g., semantic relatedness).

The third theoretical dimension is the **Information Encoding Type** of the model (hereafter referred to as **Information Type**). This dimension determines the primitive elements or building blocks of the model, and what relations between those primitive elements are encoded. Models have varied widely in this regard, in terms of the primitive elements of the model, including words, documents/discourses, objects and events in the world, semantic features, and specific kinds of semantic relations and roles. Likewise, the relations between those elements have varied widely, including a priori (theory-based) linkages, normative association strength, measures of association or co-occurrence in the environment, and similarity. Often it is argued that this dimension is the most distinguishing and critical aspect of a model, but model comparisons rarely vary only this dimension while holding all others constant.

Table 1. *Examples of Published Semantic Models, Classified on Six Theoretical Dimensions*

Classical Models	Information Source	Representation Structure	Encoding Type	Abstraction Mechanism	Learning Mechanism	Relatedness Measure
Semantic Feature Model Smith et al., 1974	Hand-labeled Relations	Vector Space	Concepts Defined as Feature Vectors	Predefined Abstract Concepts	Unspecified	Vector Distance
Hierarchical Semantic Network Collins & Quillian, 1969	Hand-labeled Relations	Multiple Relation-Labeled Connected Graph	Binary Predicate-Argument Relations	Predefined Abstract Concepts	Unspecified	Distance in Graph
Spreading Activation Network Collins & Loftus, 1975	Hand-labeled Relations	Unlabeled/LabeledConnected Graphs	Word Forms and Concepts in Separate Graphs	Predefined Abstract Concepts	Unspecified	Spreading Activation
Spreading Activation Network Anderson, 1983	Hand-labeled Relations	Multiple Relation-Labeled Connected Graph	Predicate-Argument Relations	Predefined Abstract Concepts	Unspecified	Spreading Activation
Wordnet Miller, 1995	Hand-labeled Relations	Taxonomy Connected Graph	Super/Subordinate Category Relations	Predefined Abstract Concepts	Unspecified	Connection vs. No Connection
Distributed Feature Models						
Semantic Relation Network Rumelhart & Todd, 1993 Rogers & McClelland, 2004	Hand-labeled Relations	Connected Graph	Concept+Relation to Feature Associations	Latent Variables (Hidden Units)	Error-driven Association Learning	Feature Activation Given Concept +Relation
Semantic Feature Network Hinton & Shallice, 1991 Plaut & Booth, 2000	Hand-labeled Relations	Connected Graph	Concept-Feature Assoc. Strength	Latent Variables (Hidden Units)	Error-driven Association Learning	Network Settling Time
Semantic Feature Network McRae et al., 1997 Cree et al., 1999	Normative Relations	Connected Graph	Concept-Feature Assoc. Strength	Latent Variables (Hidden Units)	Error-driven Association Learning	Network Settling Time
LISA Hummel & Holyoak, 2003	Hand-labeled Relations	Connected Graph	Relational Associations	Predefined Abstract Concepts	Hebbian Learning	Network Settling Time
Distributional Models						
Latent Semantic Analysis Landauer & Dumais, 1997	Linguistic Corpora	Vector Space	Log Entropy Word Doc. Co-occur.	Latent Variables Derived via SVD	Word-Document Frequency Counting	Vector Cosine
Hyperspace Analogue to Language (HAL) Lund & Burgess, 1996	Linguistic Corpora	Vector Space	Word Co-occur. Probability	None	Word-Word Co-occurrence Counting	Vector Distance
Naïve Discrimination Learning Baayen et al., 2019	Linguistic Corpora	Vector Space	Word Association Strength	None	Rescorla-Wagner Model	Vector Correlation
BEAGLE Jones & Mewhort, 2007	Linguistic Corpora	Vector Space	Word-Word Co-occurrence	None	Random Vector Accumulation	Vector Cosine
Probabilistic Topic Model	Linguistic	Vector Space	Word-Document	Latent Variables	Word-	Vector

Blei & Jordan., 2004 Griffiths et al., 2007	Corpora		Co-occurrence	(Topics) Derived via LDA	Document-Topic MCMC Sampling	Inner Product
Word2Vec Mikolov et al., 2013	Linguistic Corpora	Vector Space	Word-Context Predictability	Latent Variables (Hidden Units)	Error-driven Word Context Prediction	Vector Cosine
GloVe Pennington et al., 2016	Linguistic Corpora	Vector Space	Word-Word Co-occurrence	Latent Variables (Hidden Units)	Error-driven Co-occur. Ratio Prediction	Vector Cosine
Simple Recurrent Network Elman, 1990 Huebner & Willits, 2018	Linguistic Corpora	Vector Space	Word-Word Predictability	Latent Variables (Hidden Units)	Error-driven Learning via Prediction	MDS / Vector Correlation
GPT-2, 3, 4 Brown et al., 2020 Radford et al., 2019	Linguistic Corpora	Connected Graph	Sequence Predictability	Latent Variables (Hidden Units)	Error-driven Learning via Prediction & RLHF	Output Sequence Activation Given Input Sequence

The fourth theoretical dimension is whether the model includes a proposed **Abstraction Mechanism**, and if so, the nature of that mechanism. Here, abstraction refers to the process of extracting representative information or dimensions from the primitive elements and relations, by which the representational structure becomes more concise at the cost of details. Some earlier models, such as traditional semantic relation networks, and connectionist models employing abstract semantic features, relations, and roles, have used predefined abstract features, without specifying the mechanisms by which these abstractions arise. Sometimes this is an explicit nativist claim (as in the labeled edges in Anderson & Bower, 1974); but in most cases model makers have included abstract features without theoretical commitments to their origin. Most models, however, have some form of an abstraction mechanism whereby abstract features can be learned or inferred from the data. Amongst others, the mechanisms that have been used for this purpose include hidden layers in neural networks, latent dimensions identified by SVD, or related processes like Latent Dirichlet Allocation (Blei, Ng, & Jordan, 2003) or random vector accumulation (Jones, Kinstch, & Mewhort, 2006). A few models explicitly posit no abstraction mechanism (Baayen et al., 2016; Lund & Burgess, 1996).

The fifth theoretical dimension is whether the model includes a proposed **Learning Mechanism** concerning how the semantic knowledge and information is acquired, and if so, the nature of that learning mechanism. Early semantic memory models tended to focus on the structure of semantic memory and proposed no explicit learning mechanism to explain how semantic structure is acquired. Connectionist and neural network models almost always include a learning mechanism, one involving either unsupervised Hebbian learning, or supervised error-driven learning. Learning in distributional semantic models is often based on self-supervised prediction and error-backpropagation (in the case of neural network models), or on counting word co-occurrences and explicit matrix factorization.

The sixth theoretical dimension is the model's **Relatedness Measure**. How is relatedness between words and/or concepts operationalized? This property of a model is not independent of the others, in particular a model's Representational Structure. There are ways of measuring relatedness that work in a space but not a graph, and vice versa. Spatial models all use some way of comparing vectors in vector space, such as the distance between points or the cosine of the angle between the vectors. Graphical models usually use graphical distance or some type of spreading activation algorithm, either of the classical or connectionist variety. Some network models, usually described as "attractor" models, have used other measures of the model's behavior, such as the amount of time the model takes to "settle" into a stable state where activations are no longer changing (Cree et al., 1999; Plaut & Booth, 2000).

These six dimensions capture a tremendous amount of the theoretical variation in semantic models that have been proposed. In principle, the theoretical choices in one dimension are largely independent of the choices in the others (with the noted exception of the Relatedness Measure's dependence on the Representational Structure). For example, graphical and spatial data structures are both compatible with using linguistic, nonlinguistic, or normative data. Both data structures are compatible with a wide range of Information Types that can be represented, and whether and how learning and abstraction occur. However, despite this potential independence, choices on these dimensions tend to be very correlated in practice. This makes it hard to determine what makes one model outperform another, as they tend to vary along multiple correlated dimensions. To properly assess the theoretical questions involving single dimensions requires controlling the influence of other theoretical choices and isolating the dimension of interest.

In this work, we attempt this endeavor. We are interested in examining the effects of two of these dimensions, namely Representational Structure (space vs. graph), and Encoding Type (word co-occurrences vs. similarity based on word co-occurrence). Thus, we create four main classes of models: a co-occurrence space model, a similarity space model, a co-occurrence graph model, and a similarity graph model. Our motivations for systematically exploring these two dimensions are both practical and theoretical. From a practical standpoint, graphical models are under-represented in contemporary semantic modeling, especially in research on

automated information extraction from large naturalistic language data. Spatial models predominate in this area due to the simplicity and efficiency of algorithms used for training on large, unstructured datasets. This state of affairs exists primarily due to practical reasons (e.g., availability of efficient algorithms for training and/or inference) but does not reflect theoretical or empirical advantages of spatial over graphical models. As such, the paucity of graphical models is potentially concealing unknown benefits that would result if they could be made to perform as efficiently as contemporary spatial models. Specifically, graphical models are rarely trained on or constructed with word co-occurrence data, and even less rarely using large corpora of natural language. To address this paucity of research, we trained and evaluated a graphical model on distributional linguistic data (word co-occurrence data). In the upcoming sections, we discuss theoretical considerations for our proposal. We will argue that graphical models, while under-represented in the literature, may be useful for addressing limitations on the representational abilities of contemporary spatial models.

A Limitation of Spatial Models

All spatial models represent words as vectors, whose dimensions may be populated with features specified directly by the modeler, from norming studies, or from naturalistic data like linguistic corpora. For example, the dimensions of a feature-based vector for *fish* might consist of the proportion of raters who judged *fish* on some feature dimension (e.g., 'can-fly', 'can-swim', 'has-beak'). Most proposed spatial models derive their semantic information from linguistic data from a naturalistic corpus, such as how often *fish* co-occurs with other words, or the number of times *fish* occurs in each of a set of documents. Most spatial models normalize these co-occurrence counts in some manner, such as converting co-occurrences to pointwise mutual information values (Bullinaria & Levy, 2007).

Given the large number of dimensions in models trained with vocabulary sizes that are often in the tens of thousands, dimensionality reduction is typically performed after vectors are populated with co-occurrence counts. Typically, this involves using an algorithm like Singular Value Decomposition (SVD, Landauer & Dumais, 1997), Latent Dirichlet Allocation (LDA, Blei et al, 2003), or Random Vector Accumulation (RVA, Jones & Mewhort, 2007). In addition to reducing the size of the vectors, dimensionality reduction also serves several other useful purposes. These procedures reduce the sparsity of semantic vectors (which, if unreduced, typically contain mostly zeros). Another consequence is that dimensionality reduction serves as a method for generating more abstract representations, since the resulting dimensions serve as latent variables that aggregate information in multiple rows or columns with similar covariance. For example, in a co-occurrence matrix where *robin*, *eagle*, and *crow* all co-occur with *wing*, *fly*, *feathers*, and *beak*, the columns for these words can end up being combined into one or more abstract latent dimensions on which words related to birds load highly compared to other words like *airplane* or *penguin* which share fewer features with birds.

Despite their widespread use in both NLP applications and cognitive modeling, spatial models suffer important limitations with respect to accounting for the full range of human semantic abilities. One critical issue is that spatial models cannot distinguish - in a principled fashion - between different types of semantic relations (such as syntagmatic vs. paradigmatic) in the same semantic space. Vector distance in a single semantic space typically represents some combined measure of multiple different types of relations, or emphasizes one or more relation types more strongly than others. To make this point clear, consider a semantic model whose similarity scores are used to guess the right answer to a multiple-choice test, where the cue is *fast* and the choices are *speedy*, *slow*, *brown*, and *pointy*. If the question is "*What is the synonym?*", versus "*What is the antonym?*", the right answer changes, and the semantic model cannot possibly use the most similar word to get both answers correct.

For a model to make principled distinctions between different kinds of relations, it would require one vector space for each relation type - an inelegant solution, especially if the number of relations one wishes to represent is large. Defenders of spatial theories of semantic cognition could question the necessity of principled distinctions between different types of lexical relations. However, their psychological reality is supported by the demonstrations that humans represent syntagmatic and paradigmatic relations and use them to construct indirect semantic relations - a signature of human cognition (Balota & Lorch 1986; McNamara & Altarriba, 1988; Chwilla & Kolk 2002).

The idea that spatial models struggle to distinguish different types of relations and to form indirect relations based on those distinctions is supported by their poor performance on tasks that require indirect relations. For example, Peterson, Chen & Griffiths (2020) examined the performance of spatial models (using Word2Vec and GloVe) on a relational analogy task, of the form *king*:*man* :: *queen*:*woman*. This type of evaluation was first reported by Mikolov et al. (2013), on the basis that a model used to account for the structure of human semantic memory should be able to represent higher-order similarities, such as the similarity between *king-man* and *queen-woman*. Peterson et al. measured the similarity between word pairs (represented as vector differences) and correlated these scores to human judgements. Consistent with the idea that spatial models cannot explicitly represent indirect relations, the authors found that the models did not perform consistently across a diverse set of analogy types. While the models successfully predicted human ratings for the relation type CASE (e.g., *soldier-gun*, *plow-earth*), they performed poorly on other relation types, such as SIMILAR (e.g., *car-auto*, *'simmer-boil'*), CONTRAST (e.g. *old-young, buy-sell*), and NON-ATTRIBUTE (e.g. *fire-cold*, *corpse-life*). Of note, the spatial models performed well with syntagmatic relations (*soldier-gun*, *plow-earth*), but poorly with paradigmatic relations (*simmer-boil, old-young*), and especially poorly with those that indirectly bind the two (e.g., *fire-cold* can be decomposed into *fire-warm* and *warm-cold)*.

It is possible that the failure of spatial models to succeed across all relational analogy types is because their representational substrate is suboptimal for flexibly combining different types of relatedness among words, and to use such combinations to infer the strength of indirectly related word pairs. Our experiments below are designed to test this limitation explicitly. Strong performance on

our selectional preference task requires 1) the ability to represent both paradigmatic and syntagmatic relatedness in the same model, and 2) leveraging both kinds of relatedness simultaneously to predict indirect relatedness. We show that spatial models tend to represent either syntagmatic or paradigmatic relatedness, and that their failure to represent both in a principled manner limits their ability to infer indirect relatedness.

A Limitation of Graphical Models

In graphical models, words correspond to nodes in a graph, and relations among words are represented as edges between nodes. One advantage of the graphical structure over spatial models is their straightforward encoding of indirect relations. For example, the indirect relation *stripe-lion* can be represented as a chain of edges that connects *stripe* to *tiger*, and *tiger* to *lion*.

While this property of graphs makes them promising for inferring indirect relations, previously proposed graphical models suffer from a limitation not shared by existing spatial models: In contrast to spatial models, existing graphical models are typically populated either with hand-specified relations (Collins and Quillian, 1969; Steyvers & Tenenbaum 2005) or with normative word association data (Steyvers & Tenenbaum, 2005; Kenett et al., 2011; Kenett et al., 2016; Kenett et al., 2017; De Deyne, Perfors & Navarro, 2016; Kumar, Balota & Steyvers, 2020). Due to the differences in the materials used as input to graphical models compared to spatial models, which are often trained on large corpora as opposed to normative association data, it is impossible to make strong conclusions about whether differences in capabilities of the two model types are due to representational structure or information source. In addition, these normatively formed semantic networks are derived from established relations in human semantic memory and are thus useful only for characterizing the end-state of semantic development, rather than the process by which the semantic network is formed. Put differently, most graphical models of semantic knowledge were developed to account for the structure of semantic memory, not its development. In this work, we are concerned with the latter: How can we build semantic networks from language input with both efficiency and developmental plausibility in mind?

The lack of contact with learning and developmental processes is a substantial issue for graphical models. The methodological gap between language input and linguistic representations must be filled for graphical models to be trained on large, naturalistic corpora. There have been few recent investigations of this issue. That said, while some work on semantic networks by Hills et al. (2010), De Deyne et al. (2016), and Rotaru et al. (2018) has examined the construction of graphs directly from corpus data, these models were not trained on the scale at which spatial models are often trained. This makes existing graphical models unsuitable for comparison with many spatial models trained on corpora consisting of many millions and even billions of words. Given the current state-of-the art, the best spatial models may outperform the best graphical models simply because they were trained on much larger corpora. Such a comparison however would be more valid if graphical models could be trained on data that is equally large as, and in a manner comparable to, the standard method for training spatial models.

Combining Spatial and Graphical Models

To summarize, while spatial models can be efficiently trained on large corpora of natural language, graphical models excel at making inferences about indirect semantic relations. A similar point was recently made by Kumar, Steyvers and Balota (2021), who wrote that:

> ...it seems most likely that modern distributional models (specifically multimodal DSMs) provide a promising account of learning meaning from natural environments, whereas semantic network accounts provide useful conceptual tools to probe these representations and the processes that operate upon these representations. (p.19)

Given the complementary strengths of the two modeling approaches, a system that can take advantage of both strengths would be an important contribution to the field of semantic modeling. Here, we argue that such an integration can be accomplished. We propose a combined model that inherits the complementary strengths of graphical and spatial models but does not suffer the limitations of the latter. Specifically, the data structure of the proposed model is graphical, but the mechanism for deriving the graphical structure is based on the same co-occurrence counting methods used to construct many spatial models from corpus data. We then test the feasibility of using co-occurrence frequency, and word similarity as the basis for connecting word nodes. The latter is the subject of prior work (Rotaru et al., 2018), and our comparison builds upon and extends this work.

Before describing the procedure used to train our proposed model, we first illustrate the way in which spatial distributional models are typically created. Given a toy corpus consisting of two sentences *'dog chase cat. cat chase mouse'*, a four-by-four matrix of bi-directional word-word co-occurrences with adjacent words (*i.e.*, with window size = 1) can be formed, shown in Figure 1a. Here, the entry (*i,j*) in the matrix is the bi-directional co-occurrence between *i*th and *j*th word. From this word co-occurrence matrix, a word similarity matrix (shown in Figure 1b) can be constructed using some pairwise comparison between pairs of row vectors (such as the cosine of the angle between vectors). Thus, within the spatial framework, we can construct two kinds of models: The relatedness of words *i* and *j* can be obtained either from the co-occurrence count in cell (*i, j*) of the co-occurrence matrix, or the similarity score in cell (*i, j*) of the similarity matrix. Henceforth, we will refer to these two types of models as Co-occurrence Space and Similarity Space models, respectively.

Figure 1. *Illustration of the Construction of the Four Model Types*

Co-occurrence

Space		dog	cat	mouse	chase
	dog	0	0	0	1
	cat	0	0	0	2
	mouse	0	0	0	1
	chase	1	2	1	0

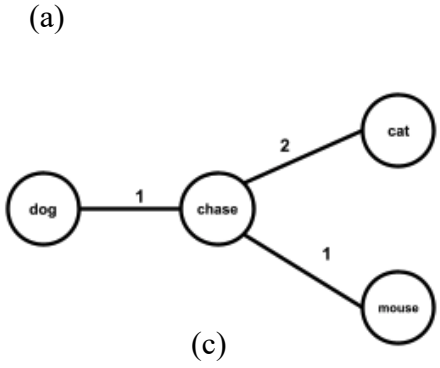
(a)

Similarity

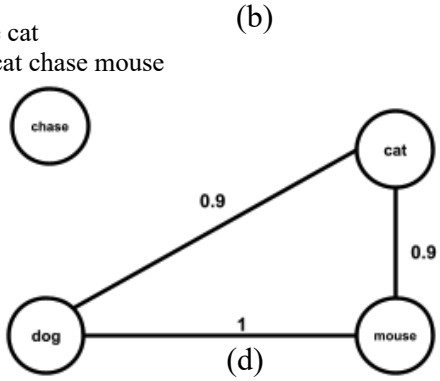
		dog	cat	mouse	chase
	dog	1	0.9	1	0
	cat	0.9	1	1	0
	mouse	1	1	1	0
	chase	0	0	0	1

(b)

Graph



(c)



(d)

Note. A schematic illustrating the construction of the four model types we investigated. The toy corpus used for model construction consists of two sentences, namely '*dog chase cat*', and '*cat chase mouse*'. **a) Co-occurrence Space.** A word co-occurrence matrix, where rows correspond to words, and columns respond to their lexical contexts. Each element represents the frequency of co-occurrence. **b) Similarity Space**. A similarity matrix derived from the co-occurrence matrix in a). Each element represents the similarity between word vectors in the co-occurrence matrix. **c) Co-occurrence Graph**. A graphical representation derived from the co-occurrence matrix. **d) Similarity Graph.** A graphical representation derived from the similarity matrix in b).

Table 2. *Model Types Investigated in the Study*

Model Types	Information Source	Representation Structure	Encoding Type	Abstraction Mechanism	Learning Mechanism	Relatedness Measure
Similarity Space	Linguistic Corpora	Vector Space	Word-word Similarity	None/SVD	Co-occurrence counting	Vector distance /Similarity/correlation
Similarity Graph	Linguistic Corpora	Connected Graph	Word-word Similarity	None/SVD	Co-occurrence counting	Spreading Activation
Co-occurrence Space	Linguistic Corpora	Vector Space	Word-word Co-occurrence	None	Co-occurrence counting	Vector distance /Similarity/correlation
Co-occurrence Graph	Linguistic Corpora	Connected Graph	Word-word Co-occurrence	None	Co-occurrence counting	Spreading Activation

Note. Models are classified on six parameter dimensions, and manipulated in two dimensions.

We can reuse the same matrices to construct corpus-derived graphical models. Because both the columns and rows refer to words we wish to represent, the co-occurrence matrix and the similarity matrix can each be considered an adjacency matrix of a graph. When deriving adjacency information from the co-occurrence matrix, the words become nodes in the graph, and are connected by an undirected weighted edge proportional to the value of the corresponding entry in the co-occurrence matrix (provided an entry is non-

zero). We refer to the resulting model as a Co-occurrence Graph (illustrated in Figure 1c). Similarly, we can construct a graphical model in which the weights of edges are derived from the similarity score of the words in the similarity matrix. We refer to the resulting model as a Similarity Graph (illustrated in Figure 1d).

In total, we consider four types of models that vary along two dimensions, namely Representational Structure (graph vs. space) and Encoding Type (co-occurrence vs. similarity) in the other. In Table 2, these four models are situated in the six-dimensional theoretical space from Table 1.

Semantic Relatedness in Graphical Structures

Like spatial models, relations represented in a semantic graph can be measured quantitatively. While relatedness in spatial models is computed by comparing vectors, relatedness in graphical models is quantified by the amount of activation that reaches a certain location in the graph. How this is computed is important. If relatedness in the graph is calculated as the edge strength between two directly connected words, then there is no difference between relatedness in the spatial and graphical models. However, graphical models often employ more sophisticated procedures, such as a spreading activation (Collins & Loftus 1975), where activation travels from one word to another along multiple weighted edges. This allows for target nodes to be activated by multiple intermediate nodes, via indirect connections. Consequently, the relatedness score computed using spreading activation can be very different from the edge weight that directly links two nodes (provided there is one). Furthermore, the relatedness score can depend strongly on the specific algorithm used to spread activation through the network. We will describe the details of the algorithm we chose in an upcoming section.

The fact that spreading activation can activate a word via multiple indirect connections has important advantages for inferring semantic relatedness. For example, De Deyne et al. (2016) showed that spreading activation processes can be used to predict human similarity judgments of weakly related word pairs and the time it takes to produce those similarity judgments. Importantly, the authors attributed the success of capturing weak relations to the spreading-activation procedure used to compute semantic relatedness. The main idea is that weakly related word pairs are linked in semantic memory only via intermediate words, and that the amount of time it takes to traverse these links can be approximated by activation spreading via indirect paths. The idea that human relatedness judgments are the result of a 'stepwise' inference procedure is consistent with the empirical literature on mediated priming effects (McNamara & Altarriba, 1988; McNamara, 1992). A second example is Rotaru et al. (2018), who implemented a spreading activation process on a graph derived from a similarity matrix built from word co-occurrence data. The authors showed that semantic inferences computed on their graphical models (using activation spreading over paths with at least two edges), provided a better fit to empirical data compared to their spatial counterparts. The data that was modeled is both accuracy and response time in lexical and semantic decision tasks, and semantic relatedness judgements tasks. Although Rotaru et al. (2018) did not explicitly analyze the role of indirect connections, they did mention that the computation of semantic relatedness involved both direct and indirect links between words. Altogether, these results demonstrate the utility of graphical structures derived from corpus data, and the potential for graphical models to provide a better account of human semantic relatedness judgments of indirectly related words compared to spatial models constructed from the same data.

To summarize, contemporary spatial and graphical semantic models are both limited as satisfactory models of semantic cognition. To overcome limitations in the representational abilities of spatial models, and the lack of developmental plausibility in the construction of graphical models, we propose an integrated approach based on 1) encoding linguistic data in a graphical data structure, and 2) a method for constructing the graph based on automatic extraction of word co-occurrence statistics from corpus data. Further, to compare our proposed integrated model with other models on our selectional preference task, we developed a procedure, based on spreading-activation, for quantifying lexical semantic relatedness in our proposed model. As all four model types are derived from the same co-occurrence matrix (Figure 1), and other modeling properties and parameters are carefully controlled, we can tease apart the contribution of the different properties of the models (graph vs. space, and co-occurrence versus similarity) to performance in our task.

We hypothesized that an integrated approach would produce more accurate semantic relatedness judgments. Our reasoning is motivated by an important difference between graphical and spatial semantic models. Specifically, we argue that, in contrast to spatial models, graphical models can simultaneously encode multiple orders of similarity (for discussion, see Shütze 1998; Artetxe et al., 2018) in the same topology. More specifically, we argue that the combination of graphical structure with a measure of relatedness based on spreading-activation, effectively combines multiple semantic similarity spaces - corresponding to successively higher levels of abstraction - in the same model. This combination makes it possible to evaluate different types of relatedness (e.g., direct syntagmatic, paradigmatic, indirect syntagmatic) in the same structure without a need to transform the underlying topology. In contrast, spatial models typically specialize in only one type of relatedness at a time, and in order to obtain measures of other kinds of relatedness, the space would need to be transformed (e.g., from a co-occurrence matrix to a similarity matrix; from a similarity matrix to a higher-order similarity matrix). We return to this point in the discussion, where we provide a more detailed explanation of the similarities and differences between spatial and graphical models. In the next section, we introduce the methods we used to train and test our models.

A World for Words

Semantic models built from linguistic corpora have several advantages. They have the practical advantage that obtaining a corpus is cheaper and easier than obtaining an equally sized normative dataset, and much easier than hand-labeling semantic relations. And perhaps most importantly, corpus-based models have a theoretical advantage, in that their structure and complexity is critical to

distribution-based theories of knowledge representation. However, semantic models built from large naturalistic corpora have a matching disadvantage: the size of the corpus, and the complexity of the information contained in it, can make it difficult to understand precisely what aspects of the input contribute to the success of a model. Popular neural network models such as GPT-3, Word2Vec, and BERT, are trained on natural language corpora containing millions or billions of words, represented across millions or even billions of parameters. This can make understanding what kind knowledge they have acquired very difficult.

There are two different strategies for researchers aiming to develop better models of human semantic knowledge. The first (and most common) approach is to focus on their fit to empirical data, such as sentence reading times, eye-tracking and EEG data obtained during sentence processing, semantic priming data, and normative judgements (of relatedness, similarity, categorization, and semantic facts). The second approach, and the one that we pursue in this paper, is to design artificial datasets that are created to highlight specific formal scenarios and can be used to test a model's formal capabilities.

Using an artificial corpus has many advantages when trying to understand the basic workings of complex distributional semantic models, such as large language models. The first and most obvious is that it allows one to precisely control the language, such that the only important sources of variability in the language are those that match the theoretical question that is being tested. This allows for control of the vocabulary size, the syntactic structure, and the semantic relationships in the language, allowing for more controlled tests of the models' abilities. A second advantage is that limiting the size and complexity of the language allows the models to be more interpretable. Instead of needing to understand and interpret billions of parameters, we can deal with models that have only dozens or hundreds. This makes understanding what the model can do much easier. There is a growing number of studies that have made use of artificial language corpora to understand the representations learned by complex models (Ars & Jones, 2017; Elman, 1990, 1991, 1993; Henighan et al., 2023; Mao, Huebner, & Willits, 2022; Perruchet & Vinter, 1998; Ravfogel, Goldberg & Linzen, 2019; Ri & Tsuruoka 2022; Rohde & Plaut, 1999; Rubin, et al., 2014; St. Clair, Monaghan, & Ramscar, 2009; Tabullo et al., 2012; Wang & Eisner, 2016; White & Cotterell, 2021; Willits, 2013).

Method

An Artificial Corpus to Describe A Simulated World

To generate the artificial language corpus used for model training, we first created a simulated artificial world, consisting of agents with goals, and events that occur as those agents pursue those goals. The events that take place in the simulated world were inspired by a highly simplified hunter-gather ecology. As the events in the simulated world unfold, linguistic descriptions were generated that narrate the events, transforming event and goal-related contingencies into word sequences. We investigate which of the computational models can learn the rules generating the contingencies in the world from the linguistic sequences.

The semantic structure of the world was governed by several interacting constraints. Agents belonged to different semantic categories, and had specific drives they were trying to satisfy (hunger, thirst and sleepiness). Agents then had action plans (consisting of sequences of specific events) that they could take to satisfy those drives. The action plans and specific actions that each agent was allowed to take were dependent on their semantic category. In addition, the allowable patients of the actions were also constrained by their semantic categories (e.g. agents of the type CARNIVORE could eat entities of the type HERBIVORE but not entities of the type FRUIT).

The categories to which the entities in the world belonged instantiated a hierarchical structure, as described in Table 3. The highest-level division was between agents, inanimate objects, and locations. Agents could be human or nonhuman animals; nonhuman animals can be carnivores or herbivores; and herbivores can be either small, medium, or large. Each category had three members.

Unlike previous works (Erk et al. 2010) that used selectional preference to evaluate semantic models trained on naturalistic corpora, we need to incorporate selectional preferences into our artificial corpus. That is, we need to manipulate which nouns can be agents or patients for each verb, and with what frequency or probability. The procedure we followed for generating the events in the world, and then for generating the corpus that describes those events, is discussed below.

Agents and Goal-Driven Event Structures

The semantic categories of an agent imposed several hierarchically structured constraints on their action plans. Agents from all categories could perform the action denoted by *drink* upon encountering entities of the type LIQUID when they were thirsty. For this action plan, the event sequence for all agents was the same. The same is true of the action plan denoted by *sleep*, which involves a single action. However, event sequences for an action plan can differ between different entities, and some are unique to a specific subset of entities. For example, only agents of the type HUMAN could perform the actions denoted by *stab* and *cook* (see Appendix A for all constraints).

The semantic categories of an agent imposed significant constraints on what they could do when they were hungry. When a member of the HERBIVORE category was hungry, it had to perform the action denoted by *search*, and when it found an entity of type PLANT, it then had to perform the actions denoted by *go_to,* and *eat*. When an entity of type CARNIVORE was hungry, it had to perform the action denoted by *search*, and when it encountered an entity of type HERBIVORE, it had to perform the actions denoted

Table 3. *Entities and Actions in the Simulated World*

Animate Agents	**Inanimate Objects and Locations**
AGENT = [HUMAN, NONHUMAN]	FOOD = [NUT, FRUIT, PLANT, AGENT]
HUMAN = [Mary, Kim]	NUT = [walnut, cashew, almond]
NONHUMAN = [CARNIVORE, HERBIVORE]	FRUIT = [apple, pear, peach]
CARNIVORE = [wolf, tiger, hyena]	PLANT = [grass, leaf, flower]
HERBIVOR = [S_HERBIVORE, M_HERBIVORE, L_HERBIVORE]	LIQUID = [water, juice, milk]
	LOCATION = [river, tent, fire]
S_HERBIVORE = [rabbit, squirrel, fox]	
M_HERBIVORE = [boar, ibex, mouflon]	
L_HERBIVORE = [bison, buffalo, auroch]	

Intransitive Action	**Transitive Actions**
rest (AGENT)	go_to (AGENT, LOCATION)
search (AGENT)	chase (AGENT, AGENT)
lay_down (AGENT)	drink (AGEN , LIQUID)
sleep (AGENT)	eat (AGENT, FOOD)
wake_up (AGENT)	reach (HERBIVORE, PLANT)
yawn (AGENT except S_HERBIVORE)	catch (CARNIVORE or HUMAN, HERBIVORE)
stretch(AGENT except S_HERBIVORE or L_HERBIVORE)	peel (HUMAN, FRUIT)
get_up (AGENT)	crack (HUMAN, NUT)
	throw_(spear)_at (HUMAN, L_HERBIVORE)
	shoot (HUMAN, M_HERBIVORE)
	trap (HUMAN, S_HERBIVORE)
	stab (HUMAN, S_HERBIVORE)
	butcher (HUMAN, NONHUMAN)
	gather (HUMAN, FOOD)
	cook (HUMAN, NONHUMAN)

Note. Upper-case words denote categories of entities, while lower-cased words denote entities or actions in the simulated world. Brackets are used to group entities that belong to the same category. Parentheses are used to group entity categories involved in action, as either agent (in first position), or patient (in second position).

by *chase*, *catch*, and *eat*. When an entity of type HUMAN was hungry, they had a wider range of possible foods, and the event structures they could use. On the one hand, they could choose the "eat_fruit" action plan, which meant they must first perform the actions denoted by *search*, *go_to*, *gather*, *peel*, and finally *eat*. On the other hand, they could choose the "eat_nuts" action plan, which consisted of the same set of actions except that a NUT requires the actions denoted by *crack* instead of *peel*. An entity of type HUMAN can also eat any member of the HERBIVORE category, but each has different sizes (marked by the prefixes S_, M_, and L_) and, consequently, require different actions to be caught and turned into food. For instance, to turn a member of M_HERBIVORE into food, an agent has to perform the actions denoted by *search*, *go_to*, *chase*, *shoot*, *catch*, *butcher*, *gather*, *'go_to fire'*, and then *cook*. To turn members of L_HERBIVORE and S_HERBIVORE into food involves a similar chain of events: To convert an entity of type L_HERBIVORE into food, an agent must perform the action denoted by *throw_(spear)_at* instead of *shoot*, and to convert an entity of type S_HERBVIORE into food, an agent must perform the actions denoted by *trap* and *stab* instead of *shoot* and *catch*.

The consequence of this causal event structure was that each category (and indeed the full hierarchical structure of the categories) was instantiated in the distributional structure of the events and their agents and patients. Animate and inanimate entities were different in a very large number of ways, particularly in the inability of inanimates to serve as agents of any events. The different inanimate entities were differentiated by the events for which they could be the patient, and by the agents that participated in those events.

Simulating Agents, Actions and Events

Using the rules that define the simulated world, we ran 10 distinct simulations, each of which differed in the random seed used to initialize the world. Each simulation included 610 animate entities in 5 categories: both members of the HUMAN category; 1 instance each of 2 randomly-selected members of the CARNIVORE category; 101 instances in each of 2 randomly-selected members of the L_HERBIVORE category; 101 instances each of 2 randomly-selected members of the M_HERBIVORE category; 101 instances of each randomly-selected members of the S_HERBIVORE category. Each simulation also included inanimate entities from

two randomly selected members of each of the NUT, FRUIT, PLANT, and LIQUID categories. Members of FRUIT, NUT, PLANT, LIQUID were resources for consumption and therefore were created as-needed so that those resources did not run out during the simulation.

At the start of each simulation, each entity was placed at a specific coordinate in the simulated world, alongside randomly placed non-agent objects (members of the FRUIT, NUT, PLANT, LIQUID, and LOCATION categories). The drives of each agent (hunger, thirst, and sleepiness) were set to random values from a range specific to the category, and were increased by a category-specific rate at each time step. In this case, the event sequences of each agents and the resultant corpus will be subjected to certain level of randomness (noises).[1] Each simulation was run for 10,000-time steps, and at each step, the action state of each agent was updated. Over the course of the simulation, if a drive crossed a critical threshold (0.8), the agent was forced to choose an action plan that would lower that drive. Some actions, like 'drink', took a fixed amount of time, while others (like 'search' and 'go_to') varied in duration. The action 'search' was variable in duration because it depended on the number of failed search trials, and the action 'go_to' depended on how far away a target was. Once an agent completed an action plan and satisfied its goal, it rested until a new drive reached the critical threshold. On average, each agent completed approximately 700 actions in a simulation.

A further constraint on the world simulation was a distinction between "resource herbivores" and "action herbivores". For instance, if a simulation included 101 instances of 'rabbit', 100 of these instances were "resource herbivores" that could participate in events initiated by agent entities. Only one 'rabbit' would be classified as an "action herbivore" that could take its own actions. This distinction was put in place to ensure that there would always be enough food resources for all entities of type HUMAN and CARNIVORE, but also to balance the number of actions that are performed by an entity of type HERBIVORE relative to other agent entities.

Generating the Corpus From Simulated Actions

Corpus generation proceeded as follows: At each time step, agent entities took turns performing actions contingent on their drive levels and the event structure in which they were situated. If an agent successfully carried out an action, a sentence describing the action was generated and added to the corpus using the formulas S_1 = Agent + IntransVerb and S_2 = Agent + TransVerb + Patient. The formula that was chosen depends on the verb type. Every sentence was followed by an optional utterance boundary marker (a minor parameter, see Table 4). We used the 10 simulations to generate 10 different corpora (the first fifty sentences in a sample corpus can be found in Appendix B), with differences that resulted from which specific members of each category were selected during random initialization of entity locations and drive values. Despite this variation, the general semantic structure of the world was extremely consistent across runs. See Appendix C for more details on the consistency of the corpora.

Table 4. *Minor Parameters Considered During Model Training*

Parameter	Options
Periods included as words in the corpus	yes, no
Co-occurrence cross sentence boundary	yes, no
Co-occurrence window size	1, 2, 7
Co-occurrence window weight	flat, inverse
Co-occurrence window direction	forward, backward, summed
Normalization	non, row-log, PPMI

Note. To mitigate the potential of values on these dimensions to influence our comparison between the two major parameter dimensions, we trained all models on all possible combinations of the six minor parameters. In total, there are 216 combinations.

On average, each corpus had 14,330 sentences and 50,204 word tokens. We experimented on different corpus sizes and decided that the current size is sufficient to encode the stipulated semantic constraints.[2] The vocabulary consisted, on average, of 10 agent nouns, 17.9 patient nouns, and 22.9 verbs of which 14.9 are transitive. For each run, we calculated the number of possible verb-noun pairs as: $N_{possible} = N_{agent} N_{verb} + N_{tv} N_{patient}$, where N_{agent} , N_{verb} , N_{tv} , $N_{patient}$, are the number of agent nouns, verbs, transitive verbs and patient nouns that occurred in a given corpus. This number indicates the number of possible events that could occur if there were no semantic constraints. However, due to the presence of constraints, only a portion of these events occurred in our simulations. On average, there were 498 possible verb-noun pairs per run, but only 42 percent of those pairs actually occurred in the corpus. It should be noted that not all inanimate entities in a simulation are necessarily described in the resulting corpus. For instance, the word

[1] We did not systematically manipulate the noises because we were primarily interested in the inferences made by the model on relations for which they had reliable input (but not perfect for most models).

[2] The corpora statistics and model performances were relatively consistent when the corpus sizes were larger. How the models would perform when given only partial information (with smaller corpus size) would be an interesting topic for future investigations.

apple may not be included in a corpus if none of the agents in a simulation performed the action 'eat' with the entity 'apple'. This flexibility resulted in different values for $N_{patient}$ and thus different $N_{possible}$ across corpora. More details can be found in Appendix D.

Table 5. *A Subset of Rules and Hypothetical Frequencies of Events*

	Rules				Hypothetical Frequencies in Corpus			
Noun	**crack**	**chase**	**eat**	**drink**	**crack**	**chase**	**eat**	**drink**
$Mary_a$	1	1	1	1	2	3	5	6
$tiger_a$	0	1	1	1	0	2	7	4
$rabbit_a$	0	0	1	1	0	0	3	3
$Mary_p$	0	1	0	0	0	6	0	0
$tiger_p$	0	0	0	0	0	0	0	0
$rabbit_p$	0	1	1	0	0	10	7	0

Note. The table demonstrates the relationship between binary rules governing what entities are allowed to perform which actions in the simulated world (left) and co-occurrence frequency in a corpus generated from that world (right). The subscripts *a* and *p* denote whether a word is the agent or patient of the verb of which it is an argument.

To illustrate the correspondence between the event semantics of the simulated world and the corpora generated from a single simulation, consider Table 5. The table shows a small subset of the binary rules governing which entities can perform which actions, and their corresponding co-occurrence frequencies in a randomly chosen corpus. For a complete specification, see Appendix A. Due to the tight coupling between the world and the corpus, the distributional statistics in the corpus can be said to be grounded in the statistics of the simulated world. Consequently, this enables us to use the generated distributional data not only for training of semantic models, but as the criterion for model evaluation.

Agents and Patients

During corpus generation, we distinguished between nouns that occurred in agent and patient position by annotating nouns with their semantic role. To highlight this fact, we use the subscripts *a* and *p* to distinguish nouns that occur in agent or patient position, respectively. This annotation approach ensures that words that occur in agent position and words that occur in patient position are disjoint. To illustrate, the event 'chase(tiger, rabbit)' was converted into a sequence of words that preserves the thematic role structure of the action, namely '$tiger_a$ *chase* $rabbit_p$'. This means that $rabbit_a$ and $rabbit_p$ are considered distinct word types in the corpus and are thus treated distinctly by each semantic model. There are two motivations for this approach. First, it allowed us to disentangle the ability of models to track thematic role assignment from their ability to learn and infer selectional preferences. Second, it allowed us to more carefully preserve the semantics of the world in the strings that are produced in the corpus. Co-occurrence does not always distinguish between thematic roles in a principled fashion, especially when word-order does not correlate with thematic role. For example, a co-occurrence-based representation of *rabbit* and *tiger* given the transitive sentence '*tiger chase rabbit*' and '*rabbit chase tiger'* will yield identical results, despite the difference in thematic role. This is problematic because a model based purely on co-occurrence may not be able to distinguish the two nouns even though each has different constraints on the kinds of actions their corresponding entities can perform in the simulated world (Table 5). See Appendix A for an example.

Experimental Design

Minor Model Parameters

While our primary goal is to compare the ability of spatial and graphical semantic models as a function of whether they are encoding co-occurrence or similarity data, we also considered the possibility that variation along other modeling parameters may influence this comparison in unforeseen ways. The reason is that there are many parameter choices known to affect the ability of distributional semantic models to perform well on various tasks, including, corpus pre-processing and the method of computing co-occurrence (Sahlgren, 2006; Bullinaria & Levy, 2007; Bullinaria & Levy, 2012). To mitigate the potential for such hidden effects, we embedded the comparison between spatial and graphical data structure within a much larger parameter space, which we refer to as 'minor' parameters. We considered six minor parameters; they are shown in Table 4. For each of the four model classes, varying along the two major parameter dimensions (graph vs. space, co-occurrence vs. similarity), we trained 216 models, to cover all combinations of minor parameters.

In addition to the major and minor parameters, we varied training of similarity models along two additional dimensions, shown in Table 6. First, we varied whether the co-occurrence matrix was reduced via SVD or left in its original form before similarities were computed. Second, we varied the metric used to compute vector similarity (distance, cosine, or correlation). As a result, there were far fewer co-occurrence models than similarity models. While this extension resulted in no additional conditions for co-occurrence models, this extension resulted in 6 additional conditions for each similarity model - beyond data structure (e.g. graph vs. space) and the 6 minor parameter dimensions. While not the primary interest of this work, we included these conditions to bolster

our ability to make strong conclusions and detect unforeseen interactions. Further details of these additional parameters can be found in the supplementary materials. [3]For more information about the effects of minor parameters on the performance of spatial models, see Sahlgren (2006), Bullinaria & Levy (2007); Bullinaria & Levy (2012), and Rubin et al., (2014).

Table 6. *Number of Models in Each of the Major Parameter Conditions*

	Encoding Type						
		Unreduced Similarity (no SVD)			Reduced Similarity (SVD)		
Data Structure	**Co-occurrence**	**Distance**	**Cosine**	**Correlation**	**Distance**	**Cosine**	**Correlation**
Spatial	n=216	n=216	n=216	n=216	n=216	n=216	n=216
Graphical	n=216	n=216	n=216	n=216	n=216	n=216	n=216

Note. For each data structure, there are 7 encoding types, i.e., the co-occurrence encoding and six similarity encodings based on different methods for computing similarity.

The complete design is thus composed of 6 minor parameter dimensions, 2 major parameter dimensions, and 2 additional dimensions that extends the number of conditions per similarity model from 1 to 6. The total number of trained models can be calculated as follows: With the extension of the 2x2 major parameter space by the two additional dimensions for similarity models, there are 2x7=14 major conditions. This is best illustrated in Table 6. In each major condition, we trained models on all possible combinations of minor parameters, namely 2x2x3x2x3x3=216. Multiplying 14 by 216 results in 3024, the number of total models trained per corpus. Training was repeated 10 times, once for each of the 10 corpora generated from different randomized runs of the simulated world.

Model Evaluation

To provide an overview of our evaluation methods, we briefly outline our procedure for model training and evaluation. For each of the 3024 models per corpus, we derived a semantic relatedness table containing the relatedness score for every word pair in the model's vocabulary. Then we evaluated each model using our selectional preference task in which we correlated a model's semantic relatedness (SR) scores to the target relatedness scores derived directly from the corpus. We report performance as an average over all 10 model instances. In the following sections, we explain 1) how semantic relatedness was computed for the spatial and graphical models; 2) how the target relatedness, used as the criterion for scoring models, was derived from the corpus; and 3) how the model-derived relatedness scores were compared to the target relatedness scores.

Computing Semantic Relatedness for Spatial Models

For the co-occurrence space models, semantic relatedness was calculated in the following way. Relatedness between two words in the co-occurrence space at indices *i* and *j* was calculated as the simple co-occurrence value in the co-occurrence matrix (normalized or raw depending on that model's normalization parameter setting). One complication is that these co-occurrence matrices were not always symmetric. For example, for models that track co-occurrences in the forward direction only (from the word in row *i* to the subsequent word in column *j*), the cell (*i, j*) encodes how often *j* followed *i*, and the cell (*j, i*) encodes how often *i* followed *j*. As we are using these co-occurrence values to predict relatedness in ordered sentence contexts, we always used the cell that corresponded to the appropriate order given the sentence. For example, when measuring the semantic plausibility of the sentence '$Mary_a$ *trap* $rabbit_p$', we used the cell corresponding to the frequency of $rabbit_p$ occurring after $Mary_a$. Due to this asymmetry, we denote *SR*(*x, y*) as the semantic relatedness from word *x* to word *y*, in that order, i.e. sensitive to the order of *x* and *y*. In this case, *SR*($Mary_a$, $rabbit_p$) denotes the semantic relatedness from $Mary_a$ to $rabbit_p$, evaluated by the cell ($Mary_a$, $rabbit_p$) of the co-occurrence matrix. For the similarity space models, relatedness was computed by as the similarity between the row vector for word *x* with the row vector for word *y*, using the similarity metric for that model (cosine, distance, or correlation), either before or after normalization, and either before or after SVD reduction, depending on that model's parameter settings.

Computing Semantic Relatedness for Graphical Models

One simple approach for computing semantic relatedness on graphs is to use the geodesic distance (the length of the shortest path connecting two nodes), which has previously been used to predict human judgements of semantic relatedness (Kenett et al., 2017, Kumar et al. 2020). If two nodes are connected by a path of length 1, their geodesic distance is 1, otherwise it is greater than 1. While the geodesic distance is a relatively straightforward approach to computing the relatedness between words in a network, there are reasons to believe it is insufficient for capturing finer-grained aspects of semantic relatedness. For example, the geodesic distance is not sensitive to word frequency or the weights on edges between nodes. In Figure 2, both the geodesic distance for the pairs *search-tiger* and *search-go_to* are 1, but *search* and *tiger* co-occur twice, while *search* and *go_to* only once. Judging by geodesic distance alone, the two pairs would be considered equally related. Additionally, behavioral studies have shown that human semantic similarity judgments are not symmetric (Tversky, 1977), a problem which even differentially weighted edges would not solve (without having bidirectional edges between nodes that have different weights). Finally, geodesic distance produces the so-called "hub effect" in which

[3]https://github.com/UIUCLearningLanguageLab/Humans/blob/master/Supplementary%20Materials/Minor_parameters.pdf

distantly related words may be geodesically close to each other by virtue of co-occurring with the same frequent word. For example, all words preceded by the article *the*, will be no more than two steps away from each other. We evaluated all our graphical models with geodesic distance and the performance was in general very poor: The top performing model variation scored 0.589 (average on two runs), in contrast to the top model with spreading-activation scoring 0.93 over the same two runs.

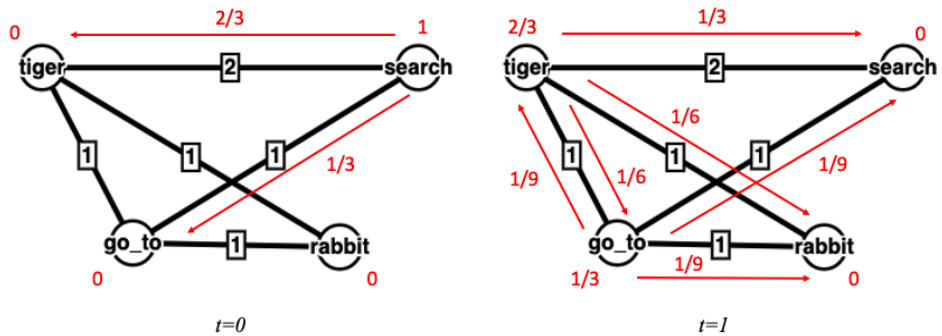
For these theoretical and practical reasons, we will measure semantic relatedness on our graphs using a spreading-activation algorithm (Anderson, 1983; Collins & Loftus, 1975). This avoids the issues mentioned above by considering both geodesic distance and co-occurrence frequency and allowing activation to spread along multiple direct and indirect pathways. To compute the semantic relatedness from word *x* to word *y* using activation-spreading, we activated *x* with strength 1, and measured the amount of activation that reached *y* at the time step at which *y* is first activated. The activation spreads through the network in the following way: At every moment, each node is activated if its activation strength is greater than zero. When a node is activated, it sends out all its activation to neighboring nodes, proportional to the weights on the edges connecting two nodes.

The process is illustrated in detail in Figure 2, in which we compute the semantic relatedness from *search* to other words in a hypothetical network. We denote $A(\text{node}_i)$ as the amount of activation of node *i*. For example, $A(tiger)=1$ means the activation on the node *tiger* is 1. At time t_0, we set $A(search)=1$ and $A(\text{node}_i)=0$, for any node *i* other than *search*. The activation of *search* will spread to its neighbors, *tiger* and *go_to*, activating each proportional to the weight of the connecting edge. By the end of t_0, $A(tiger)=2/3$ and $A(go\_to) = 1/3$ (Figure 2, left). Since this is the first-time *tiger* and *go_to* are activated, the semantic relatedness (SR) of *search* and *tiger* is 2/3, and $SR(search, go\_to) = 1/3$. At the end of t_0, *rabbit* was not activated because it was not directly connected to *search.* Therefore, in order to compute $SR(search, rabbit)$, we need to consider another time step. At the next time step t_1, the two activated words spread their activation to their neighbors. Since the connections between *go_to* and all three neighbors have the same weight, activation was spread evenly to its neighbors. Thus *search*, *rabbit* and *tiger* each received 1/9 of the activation of *go_to* (Figure 2, right). Similarly, the *A(tiger)* spreads, with amount 2/3, to its neighbors, sending 1/3 of its activation back to *search*, 1/6 to *go_to*, and 1/6 to *rabbit*. As a result, at the end of t_1, the node *rabbit* has received 1/6 activation from *tiger*, and 1/9 from *go_to*, summing up to $A(rabbit) = 5/18$. Since this is the first time step at which *rabbit* is activated, we used this number as the semantic relatedness of the pair *search-rabbit,* i.e., $SR(search, rabbit) = 5/18$. In this approach, semantic relatedness equals the amount of activation that reaches the target through the combined shortest paths to the source. For example, in Figure 2, there are two shortest paths between *search* and *rabbit* (*search-go_to-rabbit*; *search-tiger-rabbit*), and $SR(search, rabbit)$ is computed as the sum of the activation that reaches *rabbit* from both paths.

There are three factors that contribute to the magnitude of semantic relatedness computed in this way. The first is the length of the shortest paths. Activation strength diffuses as at each step when there is more than one node linked to the source node. As most nodes are likely to link to multiple nodes, less activation tends to reach the target when it is separated from the source by a longer path. The second factor is the number of shortest paths to the target**.** The larger the number of shortest paths to the target, the more activation can spread directly from the source to the target before diffusing elsewhere. The third factor is the weights of edges on the shortest paths**.** As activation diffuses proportionally to the weights, stronger weights on the shortest path will result in larger activation at the target. In our graphical models, these weights correspond to word co-occurrence or word similarity.

Figure 2:

An Illustration of Activation Spreading and Computation of Semantic Relatedness



Note. In both panels, the node search is activated with strength 1. Our annotation shows the steps involved in the computation of the semantic relatedness SR(search, tiger), SR(search, go_to), and SR(search, rabbit). Connection weights are shown in black, and activation values are shown in red. Because the example is for illustration, the thematic role subscripts of nouns are omitted.

In sum, the spreading-activation-based relatedness measure is a combination of multiple aspects of semantic networks: It is simultaneously sensitive to the geodesic distance, overall network structure, and co-occurrence frequency. Furthermore, the relatedness measure is asymmetric. In most cases $SR(x, y)$ will be different from $SR(y, x)$, and therefore has the potential to account for the asymmetry found in human judgements (Tversky, 1977; Peterson et al., 2020). Also, this approach partially avoids the hub-effect by considering the edge weights.

Note that the concept of spreading-activation on semantic networks can be traced back to classic works (Anderson, 1983; Collins & Loftus, 1975). More recent studies have implemented similar algorithms in their investigation of human semantic memory (De Deyne et al., 2016; Rotaru et al., 2018). Our spreading activation algorithm is very similar to the "random walk" algorithm used by De Deyne et al. (2016) to calculate the semantic similarity between nodes in graphs constructed from associative norming data. Both our algorithm and that of De Deyne et al. (2016) compute the relatedness of nodes that are not directly linked using the power series of the adjacency matrix. The difference is that De Deyne et al. (2016) de-emphasized the contribution of excessively long paths with a global damping parameter, while we de-emphasized the longer paths as a function of the node pair it connects. This variation is due to differences in research goals. De Deyne et al. (2016) used relatedness in their graph to predict lexical semantic similarity. However, our focus is on the relatedness between verb-noun pairs whose graphical distances vary. Thus, a global damping parameter might have unforeseen effects on our results. Instead of dampening longer paths by a global parameter as in De Deyne et al. (2016), we excluded longer paths by only including the contribution of activation via paths no longer than the distance between the source and target node. Nevertheless, our method results in a more complex algorithm compared to using the global damping parameter. Future work that tries to scale up our approach to bigger vocabulary and corpus sizes would probably be wise to adopt the global damping to reduce noise from excessively long paths.

Computing the Target Relatedness

To evaluate models, we need some way to establish the "right answer" that determines the degree to which observed and unobserved word pairs are related. There are, in principle, three ways this could be done. The first considers relatedness as a binary judgment, defined as whether two words are allowed to occur according to the rules used to constrain which actions can be performed by which entities in the simulated world. Under this definition, $Mary_a$ and *eat* would be related (because the entity denoted by *Mary* can perform the action denoted by *eat* in the simulated world), and $rabbit_a$ and *shoot* would not be related (because the entity denoted by *rabbit* cannot perform the action denoted by *shoot*). This evaluation was not our choice, as we were interested in which models could produce graded relatedness judgments that humans demonstrate based not just on whether events can occur, but on how likely they are to occur.

A second option would be to derive the correct answer from co-occurrence frequencies observed in the corpus. While this would produce graded differences, such an evaluation would not evaluate a model's ability to generalize beyond observed data, which humans are known to do. In our simulated world, 'Mary' can 'shoot' a 'boar' but cannot 'shoot' a 'rabbit' or 'water'. Despite no direct evidence, we expect that people have little trouble judging the shooting of rabbits as semantically more plausible than the shooting of water. In the real world, rabbits are more like boars than water. This is also true of the simulated world, given that the actions that 'rabbit' and 'boar' can perform overlap more than those performed by 'water' and 'boar'. Using simple co-occurrence frequencies as the gold standard against which the models are compared would not allow us to test a model's ability to make these generalizations (or, more accurately, would punish them for doing so).

To decide the "right answer" (i.e., the semantic plausibility of a noun-verb pair), we opted for a third option, namely a similarity-based procedure. In this procedure, the target scores were directly derived from the corpus co-occurrence statistics. Importantly, however, the implementation differed depending on whether a word pair is directly or indirectly related. In the former case, the target relatedness score is based on the co-occurrence frequency of a given word pair; in the latter case, the target relatedness score is based on the distributional similarity between a given noun and the nouns that co-occurred most frequently with a given verb. The higher the similarity between a noun that did not previously co-occur with a verb and nouns that did, the higher the target relatedness score. Our method of quantifying semantic relatedness is a simplified version of prior methods based on similar ideas (Erk et al., 2010)

To illustrate how we computed target relatedness, consider the relatedness between the verb *crack* and all nouns that have occurred in the corpus in agent position. The relatedness score is computed differently depending on whether a noun co-occurred with the verb or not. For nouns that co-occurred with *crack* in agent position, the semantic relatedness is simply the co-occurrence frequency. This is illustrated in Figure 3. Because $Mary_a$ co-occurred with *crack* three times, the target relatedness for the pair $Mary_a$-*crack* is 2 (Figure 3b). The computation is different for nouns that did not co-occur with a given verb in the corpus. For instance, $tiger_a$ did not co-occur with *crack* (it is not allowed to be the agent of *crack*). To quantify the relatedness of the pair $tiger_a$-*crack*, we obtained the nouns that did co-occur with *crack* in the corpus in agent position, and calculated the average of the cosine similarity between $tiger_a$ and each of those nouns. Note that the vectors used for similarity computation are the row vectors in Figure 3a. Given the example presented in Figure 3, we computed the cosine similarity between $tiger_a$ and $Mary_a$, which resulted in a value of 0.91 (Figure 3b). When there are two nouns that co-occurred with a given verb, we repeated the procedure, and averaged the resulting cosine similarities. For instance, the relatedness of $rabbit_a$-*chase* is the average of the cosine similarity of $tiger_a$-*chase* and $Mary_a$-*chase*, which ends up being 0.88.

In this way, the relatedness of a noun and verb that did not co-occur is always smaller than the relatedness of a noun and verb that did co-occur, since the co-occurrence frequency for a pair that co-occurred in the corpus is at least 1, while the cosine similarity is upper bounded by 1. In general, relatedness is highest for pairs where the noun frequently co-occurred with the verb, less high for pairs where the noun co-occurred with the verb less frequently, and lower still for pairs where the noun did not co-occur with the verb. Importantly, among the latter group of pairs, semantic relatedness is graded, such that pairs where nouns are more similar to nouns that co-occurred with the verb score higher than pairs where the noun is less similar to nouns that co-occurred with the verb.

Figure 3. *How Target and Model Relatedness Scores are Obtained*

Corpus Frequency

Thematic role	Agent			
Verb	crack	chase	eat	drink
Mary	2	3	5	6
tiger	0	2	7	4
rabbit	0	0	4	3

(a)

Target Relatedness

Thematic role	Agent			
Verb	crack	chase	eat	drink
Mary	2	3	5	6
tiger	0.91	2	7	4
rabbit	0.88	0.92	4	3

(b)

Model Relatedness

Thematic role	Agent			
Verb	crack	chase	eat	drink
Mary	0.2	0.35	0.6	0.7
tiger	0.05	0.2	0.7	0.5
rabbit	0.01	0.01	0.5	0.4

(c)

Note. This example uses only representations of agents. Panel **(a)** shows the co-occurrence frequencies for a subset of nouns and verbs in the corpus. Panel (b) shows how this co-occurrence matrix is used to compute the pairwise distributional similarities between nouns (rows) using cosine similarity. These are used to fill the cells in the co-occurrence table that have zeros. Panel **(c)** shows the model-derived semantic relatedness scores for the illustrated verb-noun pairs.

Comparing Target and Model Relatedness Scores

Figure 3 illustrates the process of how corpus-derived and model-derived relatedness scores are compared, using hypothetical data. We used Spearman correlation to evaluate the semantic relatedness scores produced by the model (Figure 3c): each column in Figure 3c is correlated with the corresponding column in Figure 3b. The resulting correlations are averaged to obtain the performance of a single model. As an example, consider the column representing the verb *crack* to agent nouns in Figure 3b and 3c, namely [2, .91, .88] and [.2, .05, .01], respectively[4]. The Spearman correlation of two column vectors is 1. The average across columns, and across all 10 corpora is the performance reported in our results. By averaging across multiple corpora, high performance is strong evidence that a model is successful at learning, representing, and inferring the fine-grained semantic relatedness between nouns and verbs.

Direct and Indirect Word Pairs

Verbs differed widely in the proportion of nouns that did and did not co-occur with them in the corpus. For instance, the verbs *eat* and *drink* co-occurred with many nouns in agent position in our artificial corpus. This is illustrated in Figure 3. The nouns *Mary*, *tiger* and *rabbit* all co-occurred with *eat* and *drink* at least once in agent position. Other verbs are more selective; for instance, verbs like *chase* and *catch* co-occurred with a smaller set of nouns in agent position. The same is true of pairs where the noun was in patient position.

Given that these two situations require different kinds of inference by the model, we split our model evaluation into two experiments. In Experiment 1, we evaluated the performance on verbs that directly co-occurred with every noun. In the Experiment 2, we evaluated performance on pairs in which the verbs co-occurred (directly) with some but not all nouns. We call the word pairs in these two experiments 'directly related stimuli' and 'indirectly related stimuli', respectively. As an illustrative example in Figure 3a, the left half, i.e. word pairs that include *crack* and *chase* are indirect stimuli, as they do not directly co-occur with all agents. On the other hand, the right half of Figure 3a, i.e. the word pairs that include *eat* and *drink* are direct stimuli. To evaluate a model, we compared each verb column in the model relatedness table to the corresponding column in the corpus-derived target relatedness table using the Spearman correlation. We averaged Spearman correlation across all direct stimuli and indirect stimuli separately, as the

[4] Each verb-column is role specific, i.e., the agent and patient nouns are under separated columns. Correlating the corresponding verb columns in the agent tables results in correlation scores of the verb to agent roles. The correlation scores of verbs to patient roles was computed separately, in the same way.

measures of a model's performance in Experiment 1 and 2, respectively.

What the average Spearman correlation tells us is how close a model's relatedness judgments are to the target relatedness. There are many alternative methods we could have chosen, and there is no straightforward single "correct" method, especially considering we cannot yet precisely formulate the computational procedure that underlies semantic inference in people. That said, we tried multiple alternatives to forming the target relatedness (the verb-noun matrix is normalized with PPMI/row-log, or no normalization; and similarity is calculated with either cosine or 2-distance), and the results did not differ qualitatively from those presented here. Further, our method closely follows a previous approach by Erk et al. (2010).

Performing Well on the Selectional Preference Task

To perform well on direct and indirect pairs requires making inferences based on two different sources of information. For direct word pairs, a model should score highest on those verb-noun combinations that occur most frequently in the corpus. For example, *Mary*$_a$ is the only agent of the verb *crack* in the corpus, and thus a model should prefer *Mary*$_a$ over other nouns for the agent role in the event *crack*. For indirect word pairs, good performance requires the ability to make inferences about unobserved noun-verb combinations, such as unobserved agents of *crack*. Although neither *tiger*$_a$ nor *rabbit*$_a$ can be the agent of the action *crack*, they should not necessarily be judged as equally implausible agents. One way to make these finer-grained judgements is to leverage indirect evidence, such as the similarity between *tiger*$_a$ and *Mary*$_a$, and *rabbit*$_a$ and *Mary*$_a$. For example, *tiger*$_a$ is more similar to *Mary*$_a$ because the entities they denote perform overlapping subsets of actions in the simulated world. Based on corpus statistics — which accurately reflect the statistics of the world — a model should infer that the entity denoted by *tiger*$_a$ is more likely to carry out an action performed by the entity denoted by *Mary*$_a$ compared to the entity denoted by *rabbit*$_a$. Notice that such an inference cannot rely on direct observation, but rather requires the integration of multiple observations, which we refer to as indirect evidence. In this example, we expect a good model to score the relatedness between *crack* and unobserved agents of *crack* based on the similarity between agents and observed agents (i.e. *Mary*$_a$*).* For example, a model should assign a higher score to the pair *tiger*$_a$*-crack* compared to *rabbit*$_a$*-crack.* Thus, good performance on our selectional preference task requires more than just tracking observed co-occurrences; instead, a model must leverage indirect evidence, namely the distributional similarity between arguments of a given verb to infer plausible, but unobserved verb-noun pairs.

Summary of the Experimental Procedure

To summarize, we performed the following steps. First, we generated an artificial corpus. Next, we computed 216 co-occurrence matrices, one for each model trained in a minor condition (Table 6). For each co-occurrence matrix, we generated six similarity matrices, and used these seven matrices as the seven spatial models. To obtain the corresponding graphical models, we used the co-occurrence or similarity matrices to create undirected graphs with edges between all nodes whose matrix values were non-zero. In cases where a matrix was not symmetric, we calculated the edge weights from the sum of the matrix with its transpose, to ensure that the resultant graph is undirected[5]. For each model, we computed the semantic relatedness score for all verb-noun pairs. This resulted in a total of $216 \times 14 = 3024$ verb-noun semantic relatedness tables, one for each model.

Next, the target relatedness scores were derived from the verb-noun co-occurrence table (Figure 3a and 3b). Then, for each model, we correlated its semantic relatedness scores for a given column in the model table with the corresponding column in the target table (Figure 3b,c) using the Spearman correlation. This resulted in a performance score for each verb in each model (separated by thematic role). These correlations were separated to distinguish direct and indirect word-pair stimuli. Finally, we averaged model scores across word pairs within each experiment (direct word pairs for Experiment 1, vs. indirect word pairs for Experiment 2). The final result was two performance scores for each model (direct and indirect respectively).

We repeated the above procedure 10 times, one for each corpus, and reported the average score for each model. All code for this paper including generation of the world and the models is available at https://github.com/UIUCLearningLanguageLab/Humans.

Hypothesis

The primary question we asked is which of the four model classes (spatial vs. graphical, crossed with co-occurrence vs. similarity) would be most successful at judging the semantic relatedness of previously observed (direct) and unobserved (indirect) syntagmatic relations. We hypothesized that graphical models should be more successful at inferring the relatedness of indirect pairs, while performing equally well at judging related pairs relative to their spatial counterparts. We discuss our reasoning in detail below.

First, our **co-occurrence space** models should perform extremely well at predicting the semantic relatedness of observed pairs, such as *Mary*$_a$*-crack*, because that is precisely what this class of models represents. More precisely, because the target semantic relatedness scores are directly derived from co-occurrence frequency, they are extremely similar to the relatedness scores produced by the co-occurrence space model. In contrast, co-occurrence space should be less successful at inferring the relatedness of indirect word pairs. The latter is true of any distributional model that has no abstraction or inference mechanism enabling integration across multiple observations.

[5] This summation will collapse a small portion of the graphical models (72 out of 1512) on the window direction dimension (forward/backward…), to be specific, the co-occurrence graphical models with no normalization.

Second, we predicted that **co-occurrence graph** models should be equally good at predicting the semantic relatedness of direct word pairs, and, importantly, are likely to be better at predicting the semantic relatedness of indirect word pairs relative to all other models. Co-occurrence graphs directly represent syntagmatic relatedness between direct word pairs (*Mary$_a$-crack*) with a direct edge between the two nodes. And while indirect word pairs (like *tiger$_a$* and *crack*) are not directly connected in the graph, they are nonetheless linked by two intermediate nodes (*tiger$_a$-chase-Mary$_a$-crack*). Combined with a spreading-activation procedure for computing semantic relatedness, this means that co-occurrence graphical models should be able to produce a non-zero relatedness score that is sensitive to the edge strength between each intermediate word and the surrounding network topology. That said, whether a particular co-occurrence graphical model will succeed in inferring the relatedness of indirect pairs is dependent on the particular combination of minor parameters (e.g. window size, normalization type). Therefore, we do not expect that all instances of the co-occurrence graph will achieve high performance. It is likely that most of the co-occurrence graphs with large window sizes or without normalization will not result in a topology useful for inference based on activation-spreading. Furthermore, we suspect that a co-occurrence graphical model with a window size of 1 should be able to capture both the direct and indirect relations better than any other model. The reason is that nouns and verbs always occur in adjacent positions in the training corpus (i.e., there are no intervening items in the word strings presented to our model). This does not mean this is the optimal window size for all tasks and learners, including humans; we return to this point in the discussion.

In sum, we predicted that there would be (1) a large amount of variation in performance due to the vast modeling space, (2) that both the spatial and graphical co-occurrence models perform equally well in predicting the relatedness of direct word pairs, and, importantly, (3) that the highest performance overall (Experiment 1 and 2) should be achieved by a restricted set of co-occurrence graph models.

Because our task requires inferences about syntagmatic relatedness, and because similarity models capture word substitutability (i.e., paradigmatic relations like *rabbit$_a$-tiger$_a$*), **similarity models** should perform overall less well than models that track co-occurrence directly. The reason is slightly different depending on whether the model is a graph or a space. As mentioned before, spatial models tend to specialize in one type of similarity, such that encoding one usually is at the cost of others. While similarity spaces might perform well at inferring the relatedness of *Mary$_a$-tiger*, this same ability will likely interfere with the model's success in predicting the relatedness of *Mary$_a$-crack*. On the other hand, while similarity graphical models are in principle able to infer different kinds of similarity, they should not be able to recover direct syntagmatic relatedness given that their computational primitive is one order of similarity above syntagmatic relatedness. Therefore, we predicted that similarity (both spatial and graphical) models will have lower performance compared to co-occurrence models when judging the relatedness of both direct and indirect word pairs.

It should be emphasized that our primary interest is not merely to identify the model that scores highest on our tasks, but rather to use performance scores as a tool for understanding the representational abilities of different types of semantic models. In particular, we were interested in which theoretical properties, and combinations thereof, enable a model to perform well. As a result, we use performance in two ways: (1) as a filter for identifying those models that warrant follow-up analyses and comparisons, and (2) to verify that our hypotheses hold up against a large amount of variation in other model parameters.

Results

Experiment 1

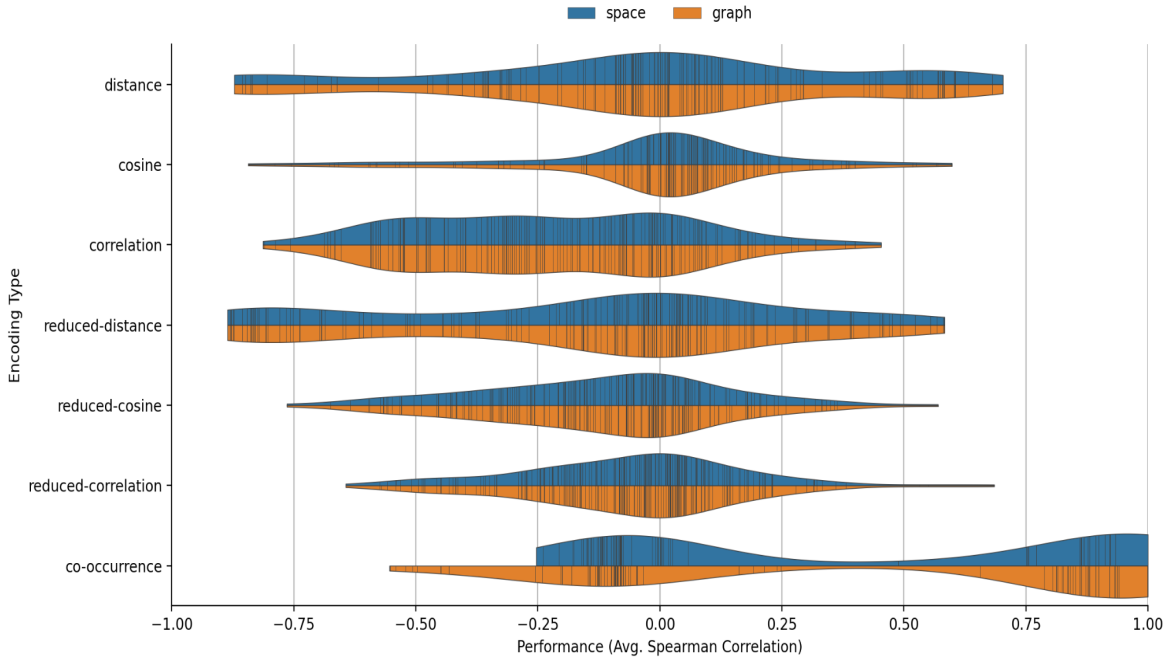
In Experiment 1, we examined the ability of models to judge the semantic plausibility of direct pairs (i.e., noun-verb pairs that occurred in the training corpus). The results are shown in Figure 4. As described in the Methods, there were two conditions varying data structure (spatial vs. graphical, shown in blue and orange in Figure 4), and seven conditions varying encoding type (shown as different violin plots in Figure 4). For each of these 14 combinations we ran 2160 models (10 randomly generated versions of the corpus for each of 216 different combinations of the minor parameters). The performance of a model averaged across all 10 corpora is shown as a line in a violin plot.[6] The thickness of a violin at a given x coordinate indicates the number of models with a performance close to that indicated by the x coordinate. The truncation at the left and right edges of a violin indicates the minimum and maximum performance of models from that group, respectively.

The figure reveals an enormous amount of variation in performance on this task. Most models resulted in a relatively poor performance, between -0.5 and +0.5. Models that surpassed +0.5 still varied along a number of parameters, such as data structure and encoding type. Notably, many similarity models in that group used distance as a similarity metric. That said, the top-performing models were those that encoded co-occurrence, and many of them achieved near perfect performance (see bottom right of Figure 4). As predicted, both spatial and graphical models are equally represented in this group of top performing models. The perfect and near-perfect performance of the co-occurrence models is not surprising. After all, the target semantic relatedness scores of direct word pairs are identical to co-occurrence frequency. The co-occurrence models therefore directly represent the information needed to perform well on this task, and do not require sophisticated inference.

---

[6] We also conducted a mixed effect model analysis using model's spearman correlation of its semantic relatedness scores with the relatedness rankings as the dependent variables, the model's Encoding Type (co-occurrence or similarity) and Representational Structure (space or graph) as predictor variables and had each run of the model as the random factor. Detailed results and analysis can be found in the supplementary materials:
https://github.com/UIUCLearningLanguageLab/Humans/blob/master/Supplementary%20Materials/2by2.pdf

Figure 4. *Average Model Performance on Selectional Preference Task for Directly Co-occurring Words*



Note. Results are broken down by data structure (graph vs. space) and type of information encoded (the seven violin plots). All information encoding types except co-occurrence reflect usage of one of the three similarity metrics, either before or after reducing the data matrix using SVD. Black lines represent average performance of 10 runs of each model with specific minor parameter settings.

Table 7. *A case study comparing the best co-occurrence and the best similarity model at judging the semantic similarity between agent nouns and the verb drink.*

Noun	target	best similarity	co-occurrence
tiger$_a$	293 (1)	.7605 (3)	.0915 (1)
wolf$_a$	289 (2)	.7597 (4)	.0912 (2)
Kim$_a$	81 (3)	.7655 (2)	.071 (3)
Mary$_a$	74 (4)	.7657 (1)	.069 (4)
squirrel$_a$	70 (5)	.7586 (7)	.068 (5)
boar$_a$	64 (6)	.7597 (4)	.0672 (6)
rabbit$_a$	63 (7)	.7575 (9)	.0669 (7)
ibex$_a$	61 (8)	.7589 (6)	.0664 (8)
buffalo$_a$	55 (9)	.7574 (10)	.065 (9)
bison$_a$	57 (10)	.7583 (8)	.059 (10)

Note. Model-derived semantic relatedness scores are shown alongside the target relatedness scores derived from the corpus. Values in parentheses are rank-transformed relatedness scores.

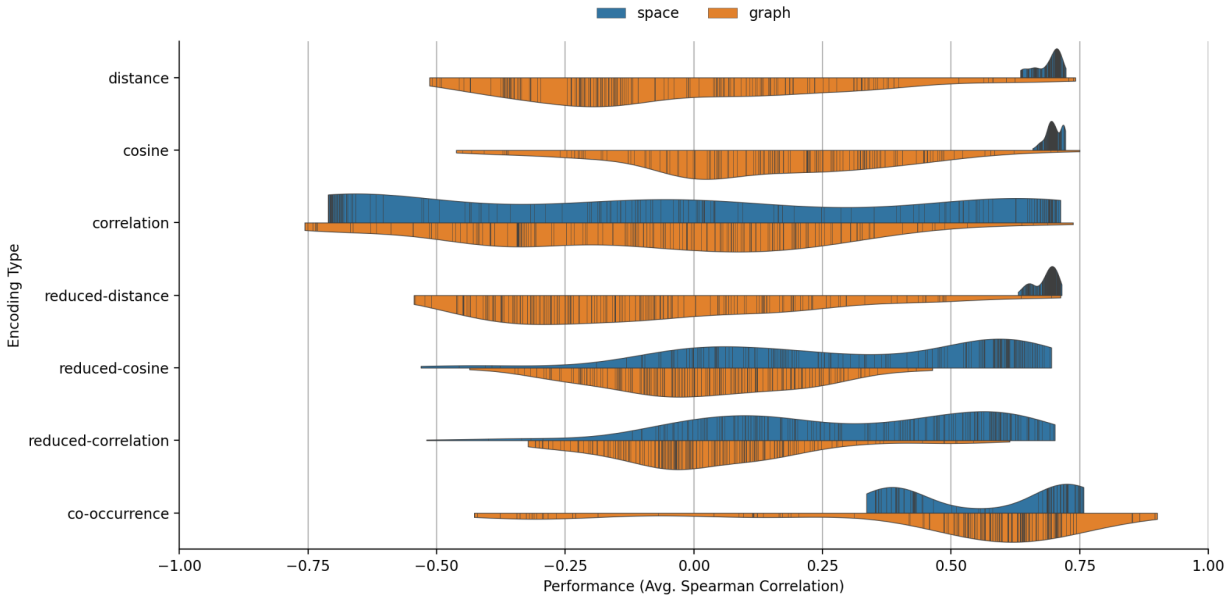
Finally, we noted that dimensionality reduction via SVD generally reduced model performance (e.g., comparing the violin plots labeled ‘distance’ and ‘reduced-distance’). There are several reasons for this: First, the raw co-occurrence count is the target semantic relatedness, so any departure (e.g., via dimensionality reduction, etc.) from co-occurrence necessarily results in worse performance unless there is a singular value encoding that co-occurrence, and only that singular value is used to calculate similarity. But because the training corpus is built from a simulated world in which all actions are diagnostic of the semantic relatedness structure among entities, the dimensionality reduction by SVD likely removed more signal than noise. To illustrate why models that encode similarity performed worse than those that encode co-occurrence, consider Table 7. In this table, we compared the best performing co-occurrence and the best performing similarity model on judging which nouns are better agents of the verb *drink* in one of the 10 corpora. In that particular corpus, the performance of the top similarity model was 0.733, while the performance of the top co-occurrence model was 1.0. This means the top similarity model did not reproduce the correct rank-ordering of observed agents of

drink according to co-occurrence frequency. The reason is that the top similarity model does not directly use co-occurrence, but, rather, must rely on word-word similarity (a transformation of co-occurrence). This transformation, as the results suggest, does not perfectly preserve co-occurrence information. It should be noted that the pattern of results presented in Table 7 is not specific to the verb *drink*, but is representative of other verbs. See Appendix E for a list of the performance of the top models on each word pair.

Experiment 2

In Experiment 1, we showed that many co-occurrence models capture the semantic relatedness of word pairs that are directly observable in the training data. Next, we compare the ability of distributional models to infer the semantic plausibility of word pairs that are *not* directly observable in the training data. In Experiment 2, we investigated the problem by evaluating models on indirect word pairs. We were particularly interested in the spatial and graphical co-occurrence models that achieved perfect performance in Experiment 1. While each was able to represent the co-occurrence pattern of observed word pairs equally well, do they differ in their ability to infer the syntagmatic relatedness between nouns and verbs that did not directly co-occur? As stated above, we predicted that the proposed co-occurrence graph models would surpass their spatial counterparts. In Experiment 2, models were evaluated on the same 10 randomly generated corpora used in Experiment 1.

Figure 5. *Average Model Performance in the Selectional Preference Task for Indirectly Related Words in Experiment 2.*



Note. Results are broken down by data structure (graph vs. space) and the type of information encoded (distance, cosine, correlation, reduced-distance, reduced-cosine, reduced-correlation, and co-occurrence). The prefix 'reduced' means that SVD was used to reduce the dimensionality of the data matrix prior to computing semantic relatedness scores. Black lines represent performance of individual models varying in minor parameter settings.

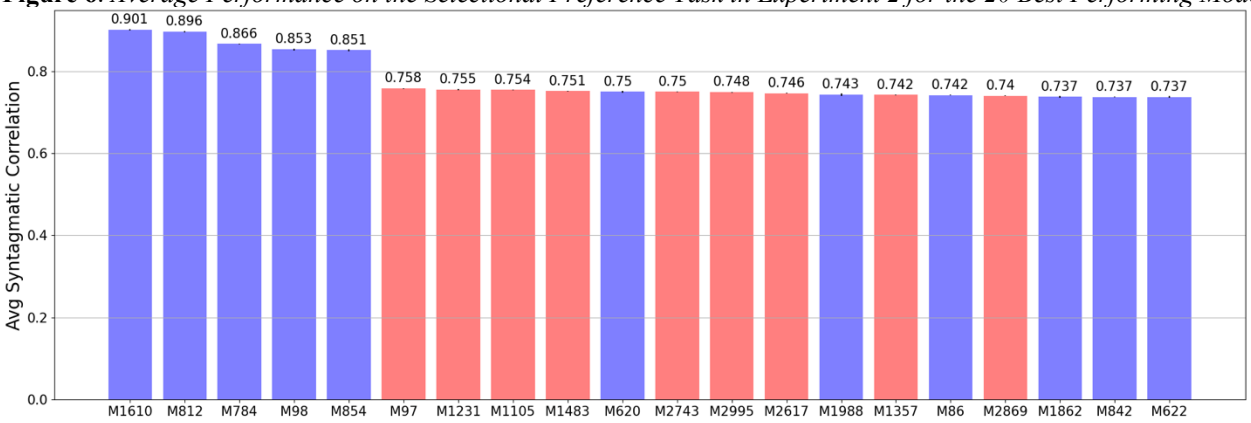
First, to get a better understanding of the overall performance across all model types, we plotted the distribution of model performance using a violin plot. The results are shown in Figure 5. As in Experiment 1, there is enormous variation in performance both within and between model types. In general, similarity space models perform relatively well on this task, while similarity graph models perform relatively poorly (many achieve an average performance below +0.5). Co-occurrence space and graph models surpass +0.5, performing better than a large proportion of other similarity models. Within that group, we found that space models clustered at approximately +0.40 and +0.75, and most graph models were more evenly distributed between 0.5 and 0.75. Furthermore, we found that there is a small minority of co-occurrence graph models that performed much better than all other models, achieving near perfect performance (the small orange hump at bottom right corner of Figure 5).

There are a number of other interesting patterns of performance in our results, including 1) the much higher degree of variability in the performance of graphical models in response to changes in the minor parameters relative to spatial models, 2) the extremely low variability and generally good performance for similarity space models without SVD and distance or cosine as the similarity metric, and 3) the generally (though not universally) worse performance of models using SVD compared to those that did not. But for the remainder of the paper, we will focus our analyses on questions related to our theoretical framework and the predictions we derived from it for those models clearly standing out from the rest in terms of combined performance in Experiment 1 and 2.

Top Performers

To determine whether, as predicted, the co-occurrence graph was in the top-performing models, we obtained the modeling parameters of the top 20 performers in Experiment 2. The average performance on indirect word pairs for the top 20 models is shown in Figure 6, and the parameters of the top 6 models are shown in Table 8. As predicted, we found that the top 5 models were co-occurrence graphs with a window size of 1. Because the same models also performed at ceiling in Experiment 1, these observations demonstrate that the proposed approach based on combining a graphical structure with co-occurrence data captured in small windows is most helpful for learning, representing, and inferring the syntagmatic relatedness of direct and indirect word pairs. These findings strengthen the claim that the proposed graphical co-occurrence model excels at (i) encoding multiple types of similarities simultaneously (e.g. syntagmatic and paradigmatic relatedness), and (ii) is able to infer the semantic relatedness of words that never co-occurred, by leveraging syntagmatic and paradigmatic relatedness in the same topology.

Figure 6. *Average Performance on the Selectional Preference Task in Experiment 2 for the 20 Best Performing Models*



Note. The top 20 models include both co-occurrence spaces (red), and co-occurrence graphs (blue). Importantly, the top-3 models are co-occurrence graphs with a window size of 1. Error bars are shown but are extremely small and difficult to see.

Lastly, these analyses revealed that, despite variation in training data (10 different runs of the simulated world, each producing a distinct corpus), and variation in minor parameters, the top performers are extremely consistent in terms of model type and performance. Not only are the standard deviations of the Spearman rank correlation miniscule for each of the best performing models, but they also differ little in their parameter setting and overall performance.

Table 8. *Average performance and parameter values for the top six models.*

Rank	Mean performance	Period	Sentence boundary	Window size	Window weight	Window type	Normalization	Encoding type	Data structure
1	0.901	no	yes	1	flat/linear	summed	log	co-occur	graph
2	0.896	yes	no	1	flat/linear	backward	log	co-occur	graph
3	0.866	yes	no	1	flat/linear	backward	ppmi	co-occur	graph
4	0.853	yes	yes	1	flat/linear	summed	log	co-occur	graph
5	0.851	yes	no	1	flat/linear	summed	log/non	co-occur	space
6	0.758	yes/no	yes	1/2/7	flat/linear	summed	log/non	co-occur	space

Note. The presence of more than one value in the same cell indicates that multiple models share the same score. In those rare cases, a change to a modeling parameter did not alter performance.

Targeted Follow-up model comparisons

What enabled the top models to perform well on the selectional preference task for indirect items? To answer this question, we compared the top co-occurrence graph model to the top co-occurrence space model. There are two reasons why such a comparison is useful. First, these two models are the best graph and space models overall. For the best performers within each level of similarity (correlation, distance, cosine, reduced and unreduced), see Appendix F. Second, these two models achieved perfect performance in Experiment 1.

To compare them, we analyzed their ability to infer the plausibility of agents for the verb *trap*. In terms of overall performance classifying plausible agents of *trap*, the top graphical model achieved a Spearman rank correlation of 0.903 between its

predicted semantic relatedness scores and the target semantic relatedness scores. In contrast, the top spatial model scored 0.701. This example is representative of differences in the performance of verbs other than *trap*.

Table 9. *A Case Study Comparing the Performance of the Best Graphical and Best Spatial Co-occurrence Model*

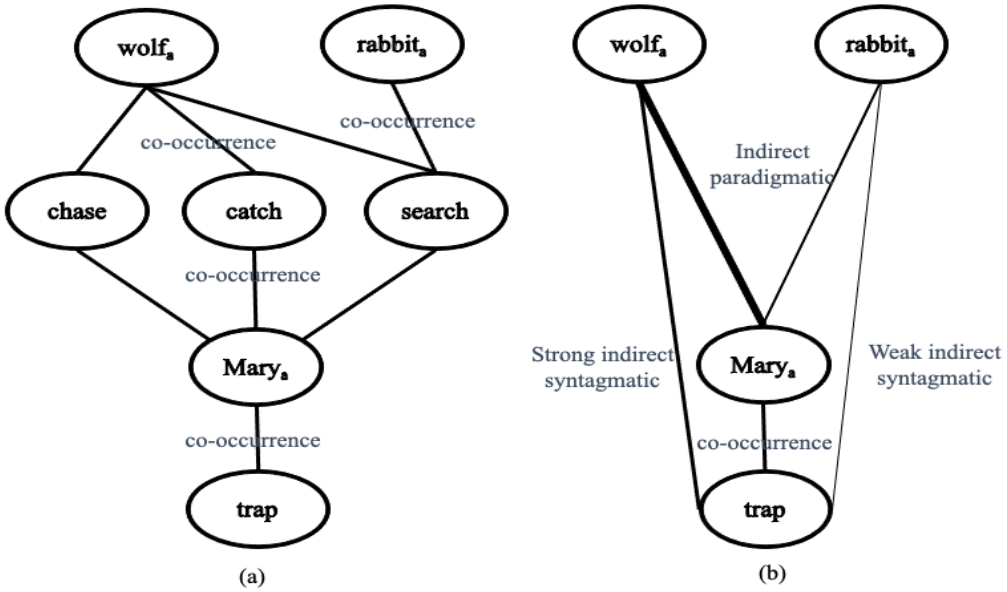
Noun	target	graphical	spatial
Kim$_a$	62 (1)	.255 (1)	.255 (1)
Mary$_a$	53 (2)	.245 (2)	.245 (2)
wolf$_a$	.83 (3)	.0553 (3)	0 (3)
tiger$_a$	.82 (4)	.0551 (4)	0 (3)
ibex$_a$	.769 (5)	.0300 (6)	0 (3)
boar$_a$	.768 (6)	.0304 (5)	0 (3)
bison$_a$	.765 (7)	.0233 (10)	0 (3)
buffalo$_a$	.764 (8)	.026 (7)	0 (3)
rabbit$_a$	.757 (9)	.0238 (9)	0 (3)
squirrel$_a$	.756 (10)	.0244 (8)	0 (3)

Note. The comparison is on the rank-ordering agents in terms of their plausibility of being an agent of the verb 'trap'. Model-derived semantic relatedness scores are shown alongside the target relatedness scores derived from the corpus. Values in parentheses are rank-transformed relatedness scores.

Looking specifically at which nouns were judged to be plausible agents (shown in Table 9), both models predicted that the best agents of the verb *trap* were nouns in the category HUMAN, which co-occur with the verb in agent position in the corpus. However, the graphical model correctly judged nouns in the categories CARNIVORE, M_HERBIVORE, S_HERBIVORE, and L_HERBIVORE to be decreasingly less plausible as an agent for *trap*. The spatial model, in contrast, did not differentiate between agent nouns in these categories. Instead, the spatial model assigned all agents that are not in the category HUMAN a relatedness of zero.

This maladaptive behavior of the co-occurrence space model can be explained in terms of how it derives semantic relatedness scores from co-occurrence data: If relatedness is derived directly from co-occurrence frequency, and co-occurrence frequency is zero, then the resulting semantic relatedness must also be zero. This cannot be remedied by tuning minor parameters, such as pre-processing or normalization. The presence of these zeros makes it impossible for co-occurrence space models to directly make fine-grained distinctions between unobserved word pairs (e.g., is *bison* or *wolf* a better agent of *trap*?).

Figure 7. *A Graphical Model Inferring the Plausibility of Unobserved Agents*



Note. The graphical model can infer that wolf$_a$ is a likely agent of trap despite having never observed wolf$_a$ as the agent of trap in the corpus it was trained on. **(a)** The nodes wolf$_a$ and Mary$_a$ are connected indirectly via three multi-edge paths that traverse nodes corresponding to the words chase, catch, and search. **(b)** The multi-edge paths are collapsed to reveal the indirect syntagmatic relationship between wolf$_a$, rabbit$_a$ and trap.

The reason for the relative success of the graphical model is that, given enough time steps, the spreading-activation algorithm produces graded relatedness scores no matter how distantly connected two nodes are. Although the node that corresponds to *trap* is not directly connected to nodes representing entities that are not of type HUMAN (i.e potential agents of *trap*), the spreading activation procedure links *trap* and potential agents via one or more indirect connections. For example, *trap* and $wolf_a$ are connected indirectly via the nodes *catch* and *chase*, (Figure 7a). By leveraging these indirect connections, the spreading activation procedure activates $wolf_a$ after three time steps, and the result is a non-zero semantic relatedness between *trap* and $wolf_a$ (Figure 7b). This can be verified by inspecting the middle column of Table 9. The graphical model correctly ranks members of CARNIVORE above HERBIVORE as agents of *trap*. The reason is that members of HUMAN are the most frequent (and only observed) agents of *trap*, and the graphical model considers entities of type CARNIVORE to be more semantically similar to entities of type HUMAN than HERBIVORE and HUMAN. This can be further explained in terms of the event semantics of the simulated world: Members of CARNIVORE (e.g. $wolf_a$) perform many of the same actions performed by members of HUMAN, and more so than HERBIVORE.

Correspondingly, members of CARNIVORE co-occur with more verbs (e.g., *catch* and *chase*) in the corpus that are shared by members of HUMAN relative to HERBIVOR. In turn, the nodes corresponding to members of CARNIVORE have a larger number of shorter — and therefore stronger — paths to nodes referring to members of HUMAN in the network. The same line of reasoning can be used to explain why the graphical model treats members of L_HERBIVORE and S_HERBIVORE as the least plausible agent of *trap*; the paths between nodes corresponding to members of L_HERBIVORE and S_HERBIVORE and nodes corresponding to members of HUMAN are fewer in number and weaker in strength.

To summarize, we have shown that graphical models tend to differentiate the plausibility of unobserved arguments of verbs, while (co-occurrence) spatial models do not. To produce graded semantic preferences, graphical models can compensate for the lack of information about unobserved co-occurrences by leveraging indirect connections via spreading activation to distantly connected nodes.

Discussion

The primary aim of this study was to compare the ability of different distributional semantic models to infer the semantic plausibility of observed and unobserved verb-noun pairs. Models were first trained on artificial corpora grounded in a simulated world with hierarchical event structures and realistic agent-environment contingencies, and then tested on a selectional preference task in two experiments. To succeed in both experiments, a model needed to encode and use fine-grained distinctions in semantic plausibility based on observed co-occurrence (i.e. direct word pairs in Experiment 1) and shared co-occurrence (i.e. indirect word pairs in Experiment 2). During evaluation, we derived semantic relatedness scores for specific verb-noun pairs from each model, and compared them against relatedness scores derived from the corpus a model was trained on. We focused our comparison on models that use distances in a vector space as a measure of relatedness, and models that use spreading activation in a graph built from the same co-occurrence data. We were also interested in the relative performance of these models as a function of whether semantic relatedness was defined in terms of word co-occurrence or word similarity. We sampled models systematically from a large space of minor parameters to better understand the contribution of individual modeling choices on downstream model performance.

Our findings can be briefly summarized as follows. First, while both graphical and spatial models performed, on average, equally well in both experiments, we found that the best graphical models performed better than the best spatial models on indirect word pairs (Experiment 2, Table 8). Second, we observed that models that used co-occurrence frequency to define its dimensions (for spatial models) or edges (for graphical models) generally produced higher and more consistent scores than those that used similarity scores derived from that co-occurrence data.

To better understand our results, we conducted targeted follow-up comparisons of the best performing models. We found that the semantic plausibility judgements produced by the best spatial model were on par with those produced by graphical models in Experiment 1 (for directly related pairs) but not in Experiment 2 (for pairs that did not directly co-occur in sentences in the corpus). Further, we found that encoding co-occurrence rather than similarity was advantageous for generalizing to indirect word pairs. The reason that the performance of spatial models lagged behind was that they assigned a semantic relatedness score of zero to word pairs that did not occur in the corpus they were trained on. In contrast, the best graphical model was able to compensate for the lack of observed co-occurrence by deriving the plausibility of an unobserved verb-noun pair via activation spreading along multiple indirect paths that link the nodes corresponding to the noun and verb in the network. The spreading activation procedure for obtaining semantic relatedness scores proved crucial, as it enabled graphical models to assign non-zero, graded semantic relatedness scores to unobserved word pairs. This means that indirect paths connecting two non-adjacent nodes appear to enable strong inferences about their semantic relatedness.

In what follows, we contextualize our findings within a principled framework for understanding the formal similarities and differences between spatial and graphical models. To preview, we suggest that semantic inference in the proposed graphical model, the co-occurrence graph, approximates a traversal of a series of increasingly higher-order similarity spaces. Our account is an attempt to pinpoint the fundamental difference between a graphical and a canonical spatial representation of distributional linguistic data, and to explain the success of the co-occurrence graph relative to its alternatives. We argue that its success critically depends on three components: 1) the graphical data structure, 2) the spreading-activation measure for computing semantic relatedness on the graph, and

3) the encoding of adjacent co-occurrence. We argue that it was the combination of these three factors that enabled the co-occurrence graphical model to succeed in our experiments.

Many Higher-Order Embedding Spaces

In order to appreciate how graphical and spatial models differ as representational substrates for semantic relatedness computations, some definitions are in order. Given a row-normalized word-by-word co-occurrence matrix C, each row corresponds to a vector representation of a target word defined as a set of co-occurrence probabilities. These word vectors reside in a multi-dimensional space, where each dimension is the (normalized) co-occurrence with a word that has appeared in the target word's context. As discussed previously, the pairwise comparison of all word vectors can be used to construct a word-to-word similarity matrix (see Figure 3). More formally, the similarity matrix is approximately the product of the co-occurrence matrix C and its transpose C^T. This is the process used to generate the similarity space models in this study. Following Schütze (1998), we refer to the vector space spanned by the co-occurrence vectors as an 'order 0' space, and the vector space spanned by the row vectors of the similarity matrix CC^T as an 'order 1' space. Vector entries in the latter space are the similarities between the vectors in the order 0 space. In this work, the co-occurrence and similarity space models are different variants of these two vector spaces.

Note that these two vector spaces are qualitatively different. Whereas entries in order 0 spaces correspond to (normalized) co-occurrence frequencies, entries in order 1 spaces correspond to the similarity between two vectors in the order 0 space. That is, in a similarity space (i.e. an order 1 space), two words are similar in terms of their pattern of co-occurrence with words in their context. While an order 0 space may capture direct syntagmatic relationships like $Mary_a$-*trap*, an order 1 space captures the paradigmatic relationship between $Mary_a$ and $tiger_a$, which share many verbs. However, as shown in the results, both types of spatial models struggle when making inferences about syntagmatically related word pairs such as $tiger_a$-*trap* that do not directly occur in the training corpus. While the verb *trap* does not directly co-occur with $tiger_a$, it does co-occur with $Mary_a$, a word that is distributionally similar to $tiger_a$. This relationship cannot be captured by relatedness computed in either an order 0 or order 1 space alone. Instead, as mentioned previously, a stepwise procedure to compute this indirect relatedness is needed, one that considers both the paradigmatic relationship between $tiger_a$ and $Mary_a$, and the syntagmatic relationship between $Mary_a$ and *trap*. One way to do this is to input vector representations from both the order 0 and order 1 spaces to the computation of relatedness. First, we can leverage the fact that the order 1 vector that corresponds to $tiger_a$ encodes the similarities between it and other words in that same space, e.g., in order to link $tiger_a$ and $Mary_a$. Second, we can quantify the strength of the relationship between $Mary_a$ and *trap* by inspecting the order 0 vector representation of $Mary_a$. If the relationship in both steps is found to be strong, this indicates that $Mary_a$ and *trap* are indirectly related.

More formally, if we have not observed word X co-occurring with word A, but we have observed Y and Z co-occurring with A, then we can infer that X should co-occur with A to the extent that it is similar to Y and Z. The relatedness of X and A can be estimated as the dot product of the order 1 vector that represents X (the row vector for X in the similarity matrix) and the order 0 vector that represents A (the row vector for A in the co-occurrence matrix). This brings us to the concept of a 'higher order' vector space. This process of creating a higher-order space from a lower order space can be continued to produce increasingly higher order spaces. Returning to the example of inferring the relatedness between X and A, we can take the order 0 vector that represents A in the co-occurrence matrix C and compute its dot product with the vector that represents X in the similarity matrix CC^T. The result can be considered a measure of the indirect relatedness between X and A.

Generalizing from vectors to vector spaces, if we take the dot product of all rows in the order 0 matrix C and the order 1 matrix CC^T, the result is the higher order similarity matrix $C(CC^T)^T$, or simply C^2C^T. In plain English, this operation involves transposing the order 1 matrix and then left multiplying it by the order 0 matrix C. In the resulting matrix, the entry at (i, j) corresponds to the dot product between the order 1 vector for word i and the order 0 vector for word j. Importantly, this process can be repeated to generate increasingly higher order vector spaces. It involves taking an existing matrix of some arbitrary order, and multiplying it by the order 0 space from which it was derived. Starting with a matrix of order 1, namely, CC^T, the process of deriving higher order spaces can be denoted by the sequence CC^T, C^2C^T, $C^2(C^T)^2$, $C^3(C^T)^2$, …, in which the two exponents are incremented in alternating fashion. Replacing the first and second exponent with the variables m and n, respectively, we can denote a space of any order using the form $C^m(C^T)^n$. An entry at (i,j) in this generalized vector space can be considered the semantic relatedness between word i and j at some abstraction level determined by m and n.

Spreading-Activation as Traversal of Increasingly Abstract Embedding Spaces

We are left with the question of how to interpret the entries in a generalized higher order vector space of the form $C^m(C^T)^n$. To interpret these higher order embedding spaces, the formalism of spreading activation in networks can be helpful. Indeed, we can show that there is a formal equivalence between the stepwise derivation of higher order spaces, and the process of spreading-activation unfolding across time steps in a graph. As an analogy to the random walk (De Deyne et al., 2016), the matrix C^m describes the activation state of the graph after m steps of activation-spreading, and the entry (i, j) of C^m is the activation arriving at node j from node i via all paths of length m. Correspondingly, the entry (i, j) in the generalized embedding matrix $C^m(C^T)^n$ is the amount of activation that 'intersects' at some intermediate locations in the graph after spreading from node i for m time steps and from node j for n time steps. It is analogous to the probability of two random walks initiated from i and j 'meeting' (intersecting) with each other on the nodes, which is exactly the entry (i, j) in the matrix $C^m(C^T)^n$. This amount of intersecting activation (random walk meeting probability) is equivalent to the similarities between the higher-order vector spaces discussed above. While this computation has been

discussed in previous work (De Deyne et al., 2016) from a methodological perspective, here, we explicitly note that this computation describes an equivalence between representing relatedness in a graph and in higher order vector spaces. A more detailed formulation of this equivalence is currently underway.

In light of this formal equivalence, we argue that there is a close correspondence between computing relatedness via spreading-activation in a graph and computing relatedness/similarity in vector spaces. In particular, we suggest that spreading-activation in the proposed co-occurrence graph can be considered as a stepwise traversal across vector spaces of different levels of abstraction (i.e. order 0, order 1, etc.) in a single topology. The benefit of the proposed model is that the same topology can be used for computing multiple orders of semantic relatedness without needing to determine when to switch to a space at a different level of abstraction. In sum, the spreading-activation procedure for computing semantic relatedness in a graph can be considered a traversal over successively higher-order vector spaces.

It should be noted that in this work, we measured activation originating at a single source node and arriving at a single target node only. That is, we set n= 0 in all our experiments. In this special case, relatedness is measured as the amount of activation intersecting at the target node, instead of at intermediate nodes distant to the target node. This method proved sufficient for the proposed model to perform well in our experiments. This makes sense considering that the two nodes that represent an indirect word pair tested in Experiment 2 were no typically no more than 2 edges (i.e. time steps) away from each other in the co-occurrence graph. The traversal of activation across the first edge corresponds to the computation of an order 1 vector representation of the source word, and the traversal of the second edge corresponds to the computation of the similarity between the order 1 vector representation of the source word and the order 0 representation of the target word.

The equivalence between the co-occurrence graph with a series of increasingly higher order vector spaces has implications for our observed differences in the behaviors of the examined models. As mentioned earlier, spatial models of the form C and CC^T only encode co-occurrence or similarity but not both. Further, the lexical relatedness derived from these models are restricted to orders 0 and 1, which do not encode information about the indirect syntagmatic relationship between $tiger_a$ and *trap*. However, a graphical model equipped with spreading activation can flexibly access lexical relatedness at multiple different levels of abstraction. For instance, the relatedness of the pair *Mary-trap* can be quantified by the activation that reaches *trap* directly from *Mary*, and the relatedness of the indirect pair $tiger_a$-*trap* can be inferred by the amount of activation that reaches *trap* after multiple time steps. Each time step of spreading-activation, therefore, corresponds to a traversal to a higher order vector space. In contrast to spatial models, the computation of these different kinds of similarities in the graph do not require moving back and forth between different kinds of representational structures, as is the case for spatial models. The ability to flexibly move between levels of abstraction should be especially useful when tasked with inferring the relatedness of words that are related as a result of multiple different kinds of distributional semantic patterns. For instance, we showed that a higher-order space is needed to capture the indirect relationship between $tiger_a$ and *trap* in our corpus, and that the co-occurrence graph was able to infer the relatedness of such word pairs. Our theoretical analysis suggests that the model was able to do so based on its ability to consult multiple orders of similarity as part of the same inference procedure (i.e. spreading-activation).

By turning our attention to the study of the relatedness between distant nodes in a graph, we are effectively studying the contribution of higher-order similarity on semantic inference. This is an important step, given that, historically, the use of strongly related words has dominated the study of the organization of semantic memory. An unintended consequence of this focus on lower-order relations has likely contributed to the success and proliferation of vector space models in accounting for psycholinguistic data. However, our work demonstrates that a graphical approach can be equally successful, and, further, may have advantages of inference on pairs of words or concepts that are less directly related. Our formal analysis is in concordance with previous empirical work which has demonstrated the psychological reality of longer paths in terms of their ability to account for additional variance in human semantic judgments beyond relatedness computed on shorter paths (De Deyne et al., 2016, Rotaru et al., 2018).

Co-Occurrence and Window Size

As we noted, we found that the best performing models not only represent lexical co-occurrence graphically, but do so using a particular encoding type, namely adjacent co-occurrences. The effect of co-occurrence encoding manifests in two ways: First, co-occurrence graphical models performed better on average relative to otherwise identical graphical models based on similarity. Second, the best performing co-occurrence graphical models were constructed by connecting nodes based only on adjacent co-occurrence (i.e. window size of 1). The advantage of encoding co-occurrence over similarity can be explained with respect to our formal analysis above: Because the similarity graphs encode similarities in CC^T directly, the order 0 co-occurrences in C ($Mary_a$-*trap*) are inaccessible, and more indirect relatedness in C^2C^T ($tiger_a$-*trap*) cannot be derived without access to the order 0 space. In contrast, using adjacent co-occurrence (order 0 similarity) as the primitive, enables the derivation of all higher order similarities. While it is straightforward to derive higher order spaces from a space of order 0, it is nearly impossible to derive lower order spaces from higher order spaces.

With the above framework in mind, it should be clearer why encoding adjacent co-occurrence was so important to the success of the best-performing models. First, note that the word pairs we used for evaluation are either agent-verb or verb-patient pairs. Importantly, agents and patients always occur next to the verbs with which they participate, meaning predicate-argument relations do not span long distances in our training corpus. Thus, a size-1 window is the optimal scope for capturing co-occurrence statistics relevant to the selectional preference tasks used in this study. Nevertheless, the choice of encoding type is a function of the semantic task and the data, and the effect of encoding type on task performance is complex (Bullinaria & Levy, 2007, 2012; Sahlgren,

2006). While we found that encoding co-occurrence in a size-1 window resulted in best performance on our task, the choice of parameters that influence how information is encoded should be empirically decided in other cases. That said, our argument from theory suggests that encoding co-occurrence is advantageous for graphical models, as activation-spreading procedure enables the derivation of higher-order relatedness from the data about order 0 relationships.

Spreading Activation as a Theory of Semantic Processing

Spreading activation need not only be considered a technical innovation for working with graphical data. Rather, our work shows that it is worth considering spreading activation as a cognitive account of the processes that underlie human performance in language tasks. Moreover, we think that processes like spreading-activation hold promise as a theory of how meaning can be computed, and not just how people search for and activate information in semantic memory. To illustrate, consider that neighboring nodes in a graph can be separated into a *structural* neighborhood defined by the network topology and a *functional* neighborhood defined by the spreading activation procedure. For example, functionally related words are those that spread activation more efficiently between each other but that are not necessarily close to each other in terms of geodesic distance. In this view, a word's meaning may be defined in terms of its functional neighborhood. Such an account proposes that the processing of a word's meaning is in part supported by the pattern of activation spreading from a given word to those words that are components of its functional neighborhood.

The idea of deriving meaning representations in terms of vectors from semantic networks has been previously investigated in the computational linguistics literature (e.g. Chakaveh et al., 2018; Grover & Leskovec, 2016; Orhan & Tulu, 2021; Perozzi et al., 2014; Pilehvar & Navigli, 2015; for review, see Grohe, 2020). For instance, Perozzi et al. (2014) obtained vectors for each node in a graph by applying the random walk method. Their system focuses on the computation of knowledge-based semantic similarity between concepts, words, and entities for structured knowledge graphs such as WordNet. These works suggest that semantic networks can be, in principle, used to perform the same kinds of semantic tasks where spatial models have been traditionally used. More broadly, with the fine-graded measures such as the spreading-activation approach proposed in this work, the derivation of vectors for representing word meanings from semantic networks (built from knowledge graphs or from natural language corpora) could rival vectors produced from spatial models in the future, and with the added benefit of transparency.

While much work is necessary to support this proposal, we think that the combination of graphical models and spreading activation-based procedures represent a promising framework for the development of mechanistic theories of human semantic memory and language comprehension.

Limitations and Future Directions

Parameter Space

While we think that our systematic approach to model comparison was essential in our ability to draw strong conclusions about the advantage of the co-occurrence graph, we did not conduct exhaustive follow-up comparisons of the contribution of minor parameters on model performance. We are aware that there are many parameter dimensions we did not consider in our analysis. Further, while extensive, our characterization of the parameter space of semantic modeling is surely incomplete, and does not straightforwardly lend itself to novel, or integrated proposals. Many semantic models simply cannot be formulated as variations inside the parametric modeling space we described without sacrificing clarity. For instance, it is not clear how artificial neural networks, like Word2Vec, should be categorized as encoding either co-occurrence or similarity relationships.

World Simulation and Corpus

We simulated a world to generate the artificial corpus, and it should be noted that the ecological endeavor may bring in a series of factors leading to potential variations in the results. For example, there are various designs in generating events, e.g. the initial drive levels and positions of the agents, the randomness in the decisions of the agent in the 'hunting' events. In addition, the corpus size might also affect the distribution of the sentence tokens incorporating the relevant semantic constraints. These variables, if systematically manipulated, may give rise to variance in corpus structure, and in turn affect the performances of the models. While beyond the scope of this paper, how the models would perform upon above-mentioned manipulations and what inference can be drawn from the result require future scrutiny.

Thematic Roles

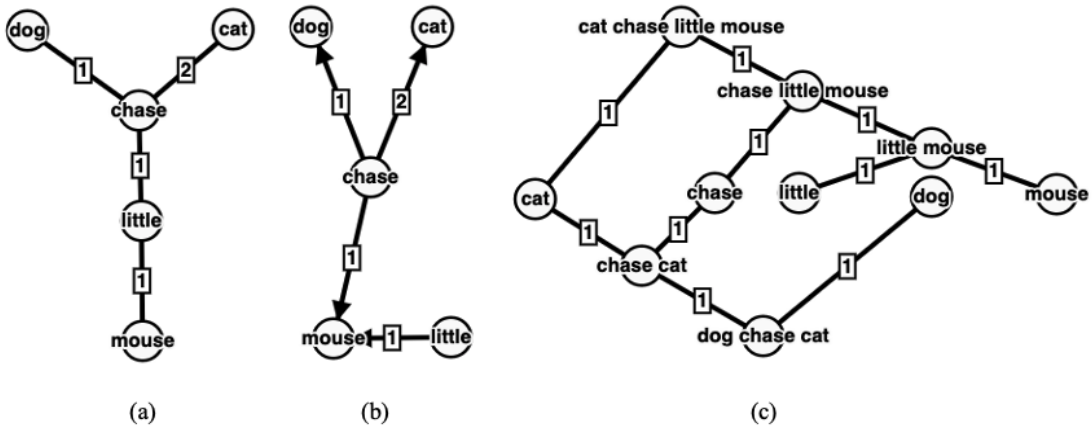
The artificial corpus used to train our models differentiates words based on their thematic role. Because human learners are not provided direct access to this kind of information, the next generation of co-occurrence graphical models should be extended with mechanisms for dealing with the potential ambiguity that results when thematic role labels are not provided. There are several ways to do so; for instance, parsing the raw language data prior to graph construction, and/or leveraging lexical statistics to infer thematic roles in a probabilistic fashion (Alishahi & Stevenson, 2008; Allen, 1997; Chang, 2004).

From Words to Phrases: Integrating Context and Syntax

There are many ways in which our artificial corpus differs from that of natural language. For instance, in contrast to the simplified input used to train our models, real-world lexical dependencies are often modulated by context information (Ferretti, McRae & Hatherell, 2001; Elman, 2009; Hare, McRae & Elman, 2003; McRae, Hare, Elman & Ferretti, 2005). As an example, the verb *carve* is compatible with many nouns, such as *knife*, and *chisel*, in instrument position, but only local linguistic (and extra-linguistic) context can determine which instrument is more plausible in each situation. To illustrate, the sentence '*Mary carves the turkey with a knife*' is more semantically plausible than '*Mary carves the turkey with a chisel*'. In addition, natural language is rife with lexical dependencies that span longer distances, such as the verb and adverb in '*Mary drinks orange juice slowly*'. Although common in natural language, neither contextual information nor long-distance dependency played a major role in the distributional semantic structure of our training corpora.

In addition to scaling the training data to more realistic language input, much more work is needed to scale the co-occurrence graph model proposed in this paper to capture more complex linguistic phenomena, such as the modulatory effect of linguistic context and long-distance dependencies. The fact that it does not do so is a symptom of a more general problem, namely that a chain built by connecting words that occur one after another in the language input cannot capture the hierarchical syntactic and semantic dependencies that exist in natural language. To handle such cases, sentences input to the co-occurrence graph need to be translated into more structured forms, such as dependency or constituent trees. An example of this procedure is shown in Figure 8. Graphical models which are built from parse trees, instead of raw co-occurrence data, organize the input in a more structured topology, and therefore should be able to capture more complex and distant dependencies. To make this possible, future work will need to generalize the spreading activation measure to operate on tree structures, so that not only lexical relatedness, but also relatedness between phrases can be computed.

Figure 8 *Semantic Networks Constructed Using Different Types of Structures.*



Note. Different sentential representations lead to distinctive structures of semantic networks: (a), dependency trees (b), or constituent trees (c) for input consisting of the sequences '*dog chase cat'* and *'cat chase little mouse*'. Whereas networks built using dependency trees and co-occurrence data are purely lexical, networks built using constituent trees encode words as well as phrases and sentences.

Modeling Work for Scaling Up

There are at least two ways in which the proposed work needs to be scaled to enable closer contact with the psychological literature. First, the model must be extended so that it can operate on more naturalistic input, beyond the simple $noun_{agent}$-verb and verb-$noun_{patient}$ constructions that characterize our artificial corpus. For instance, natural language has function words, such as determiners, conjunctions, and adpositions. Considering that many function words are extremely frequent in natural language, their inclusion could have unforeseen (potentially maladaptive) effects on the network topology, and, in turn, alter the situations in which spreading-activation can succeed. We are currently pursuing these and related questions concerning the robustness of the proposed graphical model against more realistic input. Second, the implementation of the inference procedure used to compute semantic relatedness on graphs will likely need to be made more performant to operate on much larger networks more efficiently, especially if our goal is to measure activation reaching nodes that are separated by many orders (i.e., path length). Compared to spatial models, where semantic relatedness is no more than a single vector similarity computation away, activation-spreading is a multi-step procedure with computational complexity that scales with the number of edges traversed. This is an important consideration, given that efficiency is, in practice, an important factor for researchers working with large naturalistic data. We encourage future work on more efficient implementations, both at the software and hardware level. We will continue to make our own investments in this effort, while also focusing on the theoretical issues needed to pave the way toward this goal.

Validating on Behavioral Data

This work is part of a larger and ongoing research effort that aims to bridge the gap between theoretical works in computational semantics and psycholinguistics. As a first step towards that goal, a systematic and carefully controlled comparison of model abilities is needed. This step is crucial and should take place prior to psycholinguistic experimentation so that (i) promising models can be identified, (ii) informed predictions can be generated, and (iii) the potential for unforeseen methodological issues can be reduced.

In ongoing and future work, we will extend the co-occurrence graph so that it can be trained directly on natural language input, and in the longer run, we might also integrate non-linguistic input. The current step is needed prior to comparison with human task performance, given that people's experience with language is vastly greater than that which is captured in our artificial corpora. Alternatively, we can compare relatedness judgments produced by the model and people after exposure to the same artificial input. This option, while possible in the near-term, is labor-intensive, and may not provide a complete picture of human abilities. Regardless of which alternative is chosen, we predict that human relatedness judgments for novel word pairs can be closely approximated by a graphical co-occurrence model as long as both have been provided comparable language input.

Conclusion

Using a grounded artificial corpus, and a formal model comparison framework, we systematically explored the ability of a large number of distributional semantic models to capture fine-grained and quantitative aspects of noun-verb relatedness in a novel selectional preference task. In agreement with our theoretical framing, we found that an integrated approach that combines a graphical data structure with co-occurrence information performed better than other models which utilize only graphical structure (graphical models with edges based on similarity rather than co-occurrence data), or only co-occurrence data (spatial models). More specifically, we found that strong performance on our task required a specific combination of data structure, encoding type, and co-occurrence window size. While the encoding of **adjacent co-occurrence** provides a powerful information primitive given the linguistic structure of our artificial corpora, the **graphical data structure** and the **spreading activation algorithm** of the co-occurrence graph support the access to relatedness measures of successively higher orders in the same topology. Given that models like the co-occurrence graph are relatively new in the psychological literature, much more research is needed to establish whether such systems can be used to account for a broader range of semantic tasks, and whether they are a better fit to psycholinguistic data than existing models. Leaving much of this work for future research, we are currently pursuing follow-up questions that further probe the capabilities of the proposed model, and extensions thereof in more complex tasks.

Lastly, our work also highlights the importance of teasing apart the individual contributions of different model components and their interactions to a model's success. For example, the strong performance of the proposed model was due not only to the graphical structure, but also the spreading activation-based algorithm and the encoding of co-occurrence. This highlights the importance of systematically exploring a large modeling parameter space. While doing so can yield many suboptimal configurations, it enabled us to discover rare configurations which stood out above the rest, and to better understand the full potential and limitations of our proposed model. Researchers that do not undertake extensive investigation of parameter dimensions should interpret their results with caution.

Broadly, our work is a first step towards realizing the vision outlined by Kumar et al., (2021) concerning the role of graphical structure in learning from naturalistic data:

> ...one way to reconceive semantic networks may be to use natural language (the proxy for which is typically large text corpora) and other nonlinguistic sources of information (such as images, phonology, and affective stimuli) as a starting point from which relational and concept learning occurs, and then augment this learning process with processing mechanisms that directly follow from a network-based perspective. (p.18)

In accordance with this idea, we look forward to more research on systems that operate directly on naturalistic input and encode this data in a graphical data structure.

References

Alishahi, A., & Stevenson, S. (2008). A computational model of early argument structure acquisition. Cognitive science, 32(5), 789-834.

Allen, J. (1997). Probabilistic constraints in acquisition. In A. Sorace, C. Heycock, & R. Shillcock (Eds.), *Proceedings of the GALA '97 conference on Language Acquisition* (pp. 300–305). Edinburgh, Scotland: HCRC.

Anderson, J. R., & Bower, G. H. (1974). A propositional theory of recognition memory. *Memory & Cognition, 2*(3), 406–412. https://doi.org/10.3758/BF03196896

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning & Verbal Behavior, 22*(3), 261–295. https://doi.org/10.1016/S0022-5371(83)90201-3

Baayen, R., Shaoul, C., Willits, J., & Ramscar, M. (2016). Comprehension without segmentation: a proof of concept with naive discriminative learning. *Language, Cognition and Neuroscience, 31*, 106 - 128.

Baayen, R.H., Chuang, Y., Shafaei-Bajestan, E., & Blevins, J. (2019). The Discriminative Lexicon: A Unified Computational Model for the Lexicon and Lexical Processing in Comprehension and Production Grounded Not in (De)Composition but in Linear Discriminative Learning. *Complex., 2019*, 4895891:1-4895891:39.

Asr, F.T., & Jones, M.N. (2017). An Artificial Language Evaluation of Distributional Semantic Models. *Conference on Computational Natural Language Learning*.

Artetxe, M., Labaka, G., Lopez-Gazpio, I., & Agirre, E. (2018). Uncovering Divergent Linguistic Information in Word Embeddings with Lessons for Intrinsic and Extrinsic Evaluation. *Conference on Computational Natural Language Learning*.

Balota, D. A., & Lorch, R. F. (1986). Depth of automatic spreading activation: Mediated priming effects in pronunciation but not in lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*(3), 336–345. https://doi.org/10.1037/0278-7393.12.3.336

Blei, D.M., Ng, A., & Jordan, M.I. (2003). Latent Dirichlet Allocation. *J. Mach. Learn. Res., 3*, 993-1022.

Blei, D., & Jordan, M.I. (2004). Variational methods for the Dirichlet process. *Proceedings of the twenty-first international conference on Machine learning*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language Models are Few-Shot Learners. *ArXiv, abs/2005.14165*.

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods, 39*(3), 510–526. https://doi.org/10.3758/BF03193020

Bullinaria, J. A., & Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming, and SVD. *Behavior Research Methods, 44*(3), 890–907. https://doi.org/10.3758/s13428-011-0183-8

Chakaveh, S., António, B., João, A. R., & João, S. (2018). WordNet embeddings. In Proceedings of The Third Workshop on Representation Learning for NLP (pp. 122-131).

Chang, N. (2004). Putting Meaning into Grammar Learning. *Workshop On Psycho-Computational Models Of Human Language Acquisition*.

Chwilla, D. J., & Kolk, H. H. (2002). Three-step priming in lexical decision. *Memory & cognition*, *30*(2), 217–225. https://doi.org/10.3758/bf03195282

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning & Verbal Behavior, 8*(2), 240–247. https://doi.org/10.1016/S0022-5371(69)80069-1

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82*(6), 407–428. https://doi.org/10.1037/0033-295X.82.6.407

Cree, G. S., McRae, K., & McNorgan, C. (1999). An attractor model of lexical conceptual processing: Simulating semantic priming. *Cognitive Science, 23*(3), 371–414. https://doi.org/10.1016/S0364-0213(99)00005-1

De Deyne, S., Navarro, D. J., Perfors, A., & Storms, G. (2016). Structure at every scale: A semantic network account of the similarities between unrelated concepts. *Journal of Experimental Psychology: General, 145*(9), 1228–1254. https://doi.org/10.1037/xge0000192

Deyne, S., Perfors, A., & Navarro, D. (2016). Predicting human similarity judgments with distributional models: The value of word associations. *COLING*.

Deese, J. (1962). On the structure of associative meaning. *Psychological Review, 69*(3), 161–175. https://doi.org/10.1037/h0045842

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL-HLT*.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*(2), 179–211. https://doi.org/10.1207/s15516709cog1402_1

Elman, J. (1991). Distributed Representations, Simple Recurrent Networks, and Grammatical Structure. *Mach. Learn., 7*, 195-225.

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

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive science, 33*(4), 547-582.

Erk, K., Padó, S., & Padó, U. (2010). A Flexible, Corpus-Driven Model of Regular and Inverse Selectional Preferences. *Computational Linguistics, 36*, 723-763.

Ferretti, T. R., McRae, K., & Hatherell, A. (2001). Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language, 44*(4), 516-547.

Gentner, D. (1975). Evidence for the psychological reality of semantic components: The verbs of possession. In D. A. Norman and D. E. Rumelhart (Eds.), *Explorations in Cognition* (pp. 211–246). San Francisco: W. H. Freeman.

Gershman, S., & Tenenbaum, J. (2015). Phrase similarity in humans and machines. *Cognitive Science*.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review, 114*(2), 211–244. https://doi.org/10.1037/0033-295X.114.2.211

Grohe, M. (2020). word2vec, node2vec, graph2vec, X2vec: Towards a Theory of Vector Embeddings of Structured Data. Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, 1–16. https://doi.org/10.1145/3375395.3387641

Grover, A., & Leskovec, J. (2016, August). node2vec: Scalable feature learning for networks. In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 855-864).

Hare, M., McRae, K., & Elman, J. L. (2003). Sense and structure: Meaning as a determinant of verb subcategorization preferences. *Journal of Memory and Language, 48*(2), 281-303.

Hills, T. T., Maouene, J., Riordan, B., & Smith, L. B. (2010). The associative structure of language: Contextual diversity in early word learning. *Journal of Memory and Language, 63*(3), 259–273. https://doi.org/10.1016/j.jml.2010.06.002

Hinton, G. E., & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review, 98*(1), 74–95. https://doi.org/10.1037/0033-295X.98.1.74

Huebner, P. A., & Willits, J. A. (2018). Structured semantic knowledge can emerge automatically from predicting word sequences in child-directed speech. *Frontiers in Psychology, 9,* Article 133. https://doi.org/10.3389/fpsyg.2018.00133

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review, 110*(2), 220–264. https://doi.org/10.1037/0033-295X.110.2.220

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

Jones, M. N., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language, 55*(4), 534–552. https://doi.org/10.1016/j.jml.2006.07.003

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review, 114*(1), 1–37. https://doi.org/10.1037/0033-295X.114.1.1

Jones, M., Gruenenfelder, T., & Recchia, G. (2011). In Defense of Spatial Models of Lexical Semantics. *Cognitive Science, 33*.

Kenett, Y. N., Kenett, D. Y., Ben-Jacob, E., & Faust, M. (2011). Global and local features of semantic networks: evidence from the Hebrew mental lexicon. *PloS one*, *6*(8), e23912. https://doi.org/10.1371/journal.pone.0023912

Kenett, Y. N., Beaty, R. E., Silvia, P. J., Anaki, D., & Faust, M. (2016). Structure and flexibility: Investigating the relation between the structure of the mental lexicon, fluid intelligence, and creative achievement. *Psychology of Aesthetics, Creativity, and the Arts, 10*(4), 377–388. https://doi.org/10.1037/aca0000056

Kenett, Y. N., Levi, E., Anaki, D., & Faust, M. (2017). The semantic distance task: Quantifying semantic distance with semantic network path length. *Journal of experimental psychology. Learning, memory, and cognition*, *43*(9), 1470–1489. https://doi.org/10.1037/xlm0000391

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

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*, 10.1111/tops.12548. Advance online publication. https://doi.org/10.1111/tops.12548

Lake, B. M., & Murphy, G. L. (2021). Word meaning in minds and machines. Psychological review.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211–240. https://doi.org/10.1037/0033-295X.104.2.211

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*(2-3), 259–284. https://doi.org/10.1080/01638539809545028

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments & Computers, 28*(2), 203–208. https://doi.org/10.3758/BF03204766

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.

Mao, S., Huebner, P. A., & Willits, J. A. (2022). *Compositional Generalization in a Graph-based Model of Distributional Semantics*. 1993-1999. Paper presented at 44th Annual Meeting of the Cognitive Science Society: Cognitive Diversity, CogSci 2022, Toronto, Canada.

Marcus, G. (2020). The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence. *ArXiv, abs/2002.06177*.

McNamara, T. P., & Altarriba, J. (1988). Depth of spreading activation revisited: Semantic mediated priming occurs in lexical decisions. *Journal of Memory and Language, 27*(5), 545–559. https://doi.org/10.1016/0749-596X(88)90025-3

McNamara, T. P. (1992). Theories of priming: I. Associative distance and lag. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(6), 1173–1190. https://doi.org/10.1037/0278-7393.18.6.1173

McRae, K., de Sa, V. R., & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General, 126*(2), 99–130. https://doi.org/10.1037/0096-3445.126.2.99

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, *37*(4), 547–559. https://doi.org/10.3758/bf03192726

McRae, K., Hare, M., Elman, J. L., & Ferretti, T. (2005). A basis for generating expectancies for verbs from nouns. *Memory & cognition, 33*(7), 1174-1184.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *NIPS*.

Miller, G. (1995). WordNet: a lexical database for English. *Commun. ACM, 38*, 39-41.
Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments & Computers, 36*(3), 402–407. https://doi.org/10.3758/BF03195588
Orhan, U., & Tulu, C. N. (2021). A novel embedding approach to learn word vectors by weighting semantic relations: SemSpace. Expert Systems with Applications, 180, 115146.
Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning.* Univer. Illinois Press.
Pennington, Jeffrey & Socher, Richard & Manning, Christopher. (2014). Glove: Global Vectors for Word Representation. EMNLP. 14. 1532-1543. 10.3115/v1/D14-1162.
Perozzi, B., Al-Rfou, R., & Skiena, S. (2014, August). Deepwalk: Online learning of social representations. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 701-710).
Perruchet, P., & Vinter, A. (1998). PARSER: A model of word segmentation. *Journal of Memory and Language, 39*(2), 246–263. https://doi.org/10.1006/jmla.1998.2576
Peterson, J. C., Chen, D., & Griffiths, T. L. (2020). Parallelograms revisited: Exploring the limitations of vector space models for simple analogies. *Cognition*, *205*, 104440.
https://doi.org/10.1016/j.cognition.2020.104440
Pilehvar, M. T., & Navigli, R. (2015). From senses to texts: An all-in-one graph-based approach for measuring semantic similarity. Artificial Intelligence, 228, 95-128.
Plaut, D. C., & Booth, J. R. (2000). Individual and developmental differences in semantic priming: Empirical and computational support for a single-mechanism account of lexical processing. *Psychological Review, 107*(4), 786–823. https://doi.org/10.1037/0033-295X.107.4.786
Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners.
Ravfogel, S., Goldberg, Y., & Linzen, T. (2019). Studying the Inductive Biases of RNNs with Synthetic Variations of Natural Languages. *North American Chapter of the Association for Computational Linguistics*.
Ri, R., & Tsuruoka, Y. (2022). Pretraining with Artificial Language: Studying Transferable Knowledge in Language Models. *Annual Meeting of the Association for Computational Linguistics*.
Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach.* MIT Press.
Rohde, D.L., & Plaut, D.C. (1999). Language acquisition in the absence of explicit negative evidence: how important is starting small? *Cognition, 72*, 67-109.
Rohde, D. L. T. (2002). A connectionist model of sentence comprehension and production. *Dissertation Abstracts International: Section B: The Sciences and Engineering, 63*(4-B), 1931.
Rumelhart, D. E., & Todd, P. M. (1993). Learning and connectionist representations. *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience*, *2*, 3-30.
Rotaru, A. S., Vigliocco, G., & Frank, S. L. (2018). Modeling the Structure and Dynamics of Semantic Processing. *Cognitive science*, *42*(8), 2890–2917. https://doi.org/10.1111/cogs.12690
Rubin, T. N., Kievit-Kylar, B., Willits, J. A., & Jones, M. N. (2014). Organizing the space and behavior of semantic models. *CogSci ... Annual Conference of the Cognitive Science Society. Cognitive Science Society (U.S.). Conference*, *2014*, 1329–1334.
Rumelhart, D. E., & Levin, J. A. (1975). A language comprehension system. In D. A. Norman and D. E. Rumelhart (Eds.), *Explorations in cognition* (pp. 179–208).San Francisco: W. H. Freeman.
Rumelhart, D. E., & Todd, P. M. (1993). *Learning and connectionist representations.* In D. E. Meyer & S. Kornblum (Eds.), *Attention and performance 14: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience* (p. 3–30). The MIT Press.
Sahlgren, M. (2006). The Word-Space Model : Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.
Saussure, F. de, & Riedlinger, A. (1983). *Course in general linguistics*. (R. Harris, Trans., C. Bally & A. Sechehaye, Eds.) (Ser. Paperduck). Duckworth.
Schütze, H. (1998). Automatic word sense discrimination. Computational linguistics, 24(1), 97-123.
Smith, E. E., Shoben, E. J., & Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review, 81*(3), 214–241. https://doi.org/10.1037/h0036351
Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cognitive science*, *29*(1), 41–78. https://doi.org/10.1207/s15516709cog2901_3
St. Clair, M. C., Monaghan, P., & Ramscar, M. (2009). Relationships between language structure and language learning: The suffixing preference and grammatical categorization. *Cognitive Science, 33*(7), 1317–1329. https://doi.org/10.1111/j.1551-6709.2009.01065.x
Tabullo, Á.J., Arismendi, M., Wainselboim, A., Primero, G., Vernis, S., Segura, E.T., Zanutto, S., & Yorio, A. (2012). On the Learnability of Frequent and Infrequent Word Orders: An Artificial Language Learning Study. *Quarterly Journal of Experimental Psychology, 65*, 1848 - 1863.
Tversky, A. (1977). Features of similarity. *Psychological Review, 84*(4), 327–352. https://doi.org/10.1037/0033-295X.84.4.327

Wang, D., & Eisner, J. (2016). The Galactic Dependencies Treebanks: Getting More Data by Synthesizing New Languages. *Transactions of the Association for Computational Linguistics, 4*, 491-505.

White, J., & Cotterell, R. (2021). Examining the Inductive Bias of Neural Language Models with Artificial Languages. *Annual Meeting of the Association for Computational Linguistics*.

Willits, J. A., Amato, M. S., & MacDonald, M. C. (2015). Language knowledge and event knowledge in language use. *Cognitive psychology, 78*, 1-27.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review, 114*(2), 245–272. https://doi.org/10.1037/0033-295X.114.2.245

Appendix A
Argument Structure Rules

Table A1: *Rules Governing Which Entities are Allowed to Perform Which Actions*

Agent-Verb Relation

	Human	Carnivore	Herbivore1	Herbivore2	Herbivore3
rest	1	1	1	1	1
search	1	1	1	1	1
lay_down	1	1	1	1	1
sleep	1	1	1	1	1
wake_up	1	1	1	1	1
yawn	1	1	0	1	1
Stretch	1	1	0	1	0
get_up	1	1	1	1	1
go_to	1	1	1	1	1
throw_(spear)_at	1	0	0	0	0
chase	1	1	0	0	0
shoot	1	0	0	0	0
catch	1	1	0	0	0
trap	1	0	0	0	0
stab	1	0	0	0	0
gather	1	0	0	0	0
butcher	1	0	0	0	0
cook	1	0	0	0	0
eat	1	1	1	1	1
peel	1	0	0	0	0
crack	1	0	0	0	0
drink	1	1	1	1	1
reach	0	0	1	1	1

Verb-Patient Relation

	Human	Herbivore1	Herbivore2	Herbivore3	Nut	Fruit	Plant	Liquid	Location
go_to	1	1	1	1	1	1	1	0	1
throw_(spear)_at	0	0	0	1	0	0	0	0	0
chase	1	0	1	1	0	0	0	0	0
shoot	0	0	1	0	0	0	0	0	0
catch	0	1	1	1	0	0	0	0	0
trap	0	1	0	0	0	0	0	0	0
stab	0	1	0	0	0	0	0	0	0
gather	0	1	1	1	0	0	0	0	0
butcher	0	1	1	1	0	0	0	0	0
cook	0	1	1	1	0	0	0	0	0
eat	0	1	1	1	1	1	1	0	0
peel	0	0	0	0	0	1	0	0	0
crack	0	0	0	0	1	0	0	0	0
drink	0	0	0	0	0	0	0	1	0
reach	0	0	0	0	0	0	1	0	0

Note. Rules are separated by whether the entity is an agent or a patient. Columns are labeled by 5 agent noun categories for the agent-verb pairs and 9 patient noun categories for the verb-patient pairs, each of which contains three nouns with identical behavior. Rows are labeled by verbs. Only transitive verbs are included in verb-patient pairs.

Appendix B
An Example Artificial Corpus

Table B1

First 50 Sentences of a Randomly Selected Sampling Corpus

Sentence 1-10	Sentence 11-20	Sentence 21-30	Sentence 31-40	Sentence 41-50
$Mary_a$ go_to $river_p$	$rabbit_a$ go_to $river_p$	$squirrel_a$ go_to $river_p$	$boar_a$ go_to $river_p$	$ibex_a$ go_to $river_p$
Kim_a go_to $river_p$	$rabbit_a$ go_to $river_p$	$squirrel_a$ go_to $river_p$	$boar_a$ go_to $river_p$	$ibex_a$ go_to $river_p$
$wolf_a$ go_to $river_p$	$rabbit_a$ go_to $river_p$	$squirrel_a$ go_to $river_p$	$boar_a$ go_to $river_p$	$ibex_a$ go_to $river_p$
$tiger_a$ go_to $river_p$	$rabbit_a$ go_to $river_p$	$squirrel_a$ go_to $river_p$	$boar_a$ go_to $river_p$	$ibex_a$ go_to $river_p$
$rabbit_a$ go_to $river_p$	$squirrel_a$ go_to $river_p$	$boar_a$ go_to $river_p$	$ibex_a$ go_to $river_p$	$bison_a$ go_to $river_p$
$rabbit_a$ go_to $river_p$	$squirrel_a$ go_to $river_p$	$boar_a$ go_to $river_p$	ibexa go_to $river_p$	$bison_a$ go_to $river_p$
$rabbit_a$ go_to $river_p$	$squirrel_a$ go_to $river_p$	$boar_a$ go_to $river_p$	$ibex_a$ go_to $river_p$	$bison_a$ go_to $river_p$
$rabbit_a$ go_to $river_p$	$squirrel_a$ go_to $river_p$	$boar_a$ go_to $river_p$	$ibex_a$ go_to $river_p$	$bison_a$ go_to $river_p$
$rabbit_a$ go_to $river_p$	$squirrel_a$ go_to $river_p$	$boar_a$ go_to $river_p$	$ibex_a$ go_to $river_p$	$bison_a$ go_to $river_p$
$rabbit_a$ go_to $river_p$	$squirrel_a$ go_to $river_p$	$boar_a$ go_to $river_p$	$ibex_a$ go_to $river_p$	$bison_a$ go_to $river_p$

Note. The randomly sampled corpus has 17719 sentences. Nouns in sentences are marked by thematic roles. Consecutive and identical sentences occur because there are multiple entities (e.g. multiple members of $tiger_a$) engaging in identical behavior in the simulated world.

Appendix C

Consistency of the Artificial Corpora and Robustness of the Results

To show that the corpora generated for different simulations of the hunter-gatherer world are consistent in terms of their distributional semantic properties, we calculated the average pairwise correlation of the corpus-derived target relatedness scores. The result indicates that there is little variance in the distributional semantic structure of corpora across different simulations. We also conducted the same analysis on model-derived relatedness scores, and found that most models behaved consistently across different corpora. The consistency of the corpus and model predictions across runs suggest our results are robust against variance caused by the randomness in our simulations.

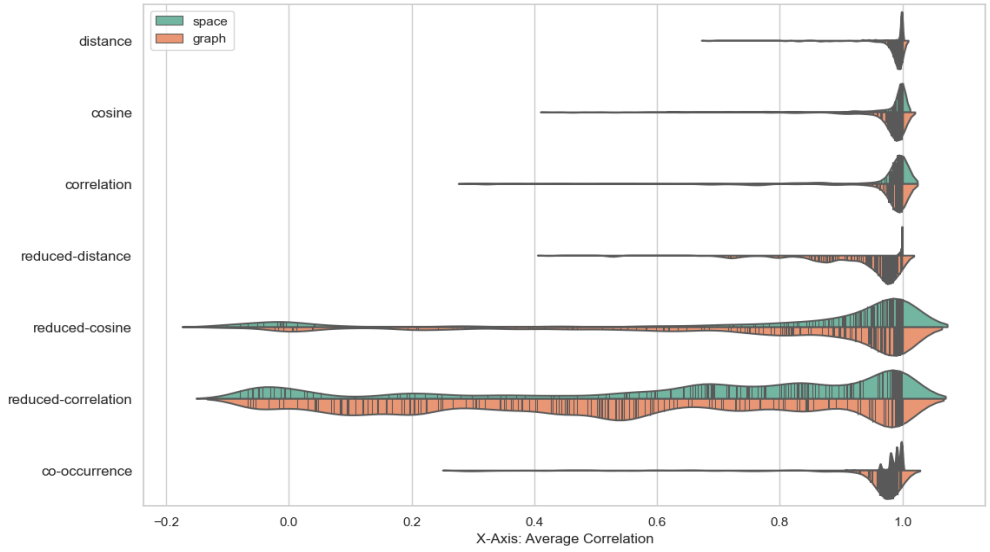
In this section, we provide more details about the methods used to produce these results. First, for each corpus, the target relatedness between each verb and noun (separated by thematic role) were grouped by the semantic category of the noun (e.g. HUMAN, CARNIVORE). We did this because the vocabularies across corpora might differ slightly. This can happen if, for instance, an inanimate entity was not engaged in any event due to the random nature of the simulation of the hunter-gatherer world. In some simulations, for instance, the word *apple* does not occur in the corpus because no apple was eaten by any agent. By collapsing nouns to their category, we guaranteed that, at the level of the noun category, there would be no missing data. As an example, if the pairs $Mary_a$*-crack* and Kim_a*-crack* are related with strength 6 and 8, respectively, the target relatedness at the level of the category HUMAN would be 7, for that corpus. In this way, we obtained the target relatedness between each verb and all 10 noun categories for all 10 corpora. A portion of the results (for corpus 1 and the verb *going_to*) are shown in Table C1. The target relatedness in each row is the average relatedness derived from the corpus, between the verb *going_to* and all members of an entity category indicated by that row.

Table C1: *Collapsed target relatedness for verb going_to in corpus 1.*

Verb	Noun category	Thematic Role	Target relatedness
going_to	HUMAN	Agent	478.5
going_to	CARNIVORE	Agent	432.5
going_to	L_HERBIVORE	Agent	92
going_to	M_HERBIVORE	Agent	125
going_to	S_HERBIVORE	Agent	133
going_to	L_HERBIVORE	Patient	49
going_to	M_HERBIVORE	Patient	53.5
going_to	S_HERBIVORE	Patient	38.5
going_to	DRINK	Patient	-1
going_to	PLANT	Patient	175
going_to	LOCATION	Patient	630
going_to	FRUIT	Patient	-.728
going_to	NUT	Patient	-.728

To quantify the extent to which target-derived relatedness vary across corpora, we performed the following steps: One corpus has 22 verbs, and the other nine have the same 23 verbs, and in every corpus the verbs collocated with the same 12 out of the 14 noun categories (5 agent noun categories and 9 patient noun categories), resulting in the identical sized verb-noun category target relatedness score vectors (22 verbs times 12 categories, ending up 264 dimensions). Finally, we computed the pairwise Pearson correlation of the 10 vectors of average relatedness scores (one for each of the 10 corpora). Of the resulting 45 correlations we found that the mean was 0.98 and the standard deviation was 0.02. This shows that the corpus-derived target relatedness scores are highly consistent across corpora.

Figure C1. Average model performance, as measured by the average Spearman rank correlation across 10 corpora.



Note. Performance is separated by Encoding Type and Data Structure. Both direct and indirect word pairs are included in this analysis.

Next, we repeated the same method for all 3024 models: For each model, the model-derived relatedness scores were collapsed by category, resulting in the same sized model relatedness vector for each corpus. We obtained 45 pairwise correlations and then averaged them. This resulted in 3024 model-specific correlations of relatedness scores averaged across the 10 corpora. We plotted in Figure C1 the 3024 data points in the same structured violin plot used in Figure 4, such that each of the 14 half violins corresponds to the 216 model variations in one of 14 conditions in Table 6. Notice that, with the exception of models in the 'reduced-correlation' condition, the average correlation is very high (above 0.9). In sum, these results show that most models produced relatively consistent inferences about verb-noun relatedness across different corpora.

Appendix D
Information of the Artificial Corpora

Table D1
Verb-noun Pair Information of the Corpora.

Corpus	Number of possible verb-noun pairs	Observed verb-noun pairs	Rate of observed pairs	Number of observed patients
1	485	206	.425	17
2	515	211	.410	19
3	515	213	.414	19
4	500	209	.418	18
5	515	212	.412	19
6	485	206	.425	17
7	460	202	.455	16
8	500	208	.416	18
9	515	212	.412	19
10	485	207	.427	17

Note. The number of possible verb-noun pairs is the total number of possible agent-verb and verb-patient pairs. The number of possible agent-verb pairs is calculated by multiplying the number of observed agents with the number of observed verbs. Similarly, the number of possible verb-patient pairs is computed by multiplying the number of observed transitive verbs with the number of observed patients. For example, in corpus 1, all models observed 10 nouns in agent position and 16 nouns in patient position. Further, given that there are 15 transitive verbs and 21 verbs in corpus 1, the total number of possible verb-noun pairs is $(10 \times 21) + (16 \times 15) = 450$.

Appendix E
Overall Analysis of Model Performances for Direct Stimuli

Table E1: *Performance of the Best Spatial and Best Graphical Model on the Selectional Task Using Direct Word Pairs in Simulation 1*

Verb-role	Co-occurrence Graph	Similarity Graph
drink-agent	1	.733
eat-agent	1	.903
get_up-agent	1	.782
go_to-agent	1	.915
lay_down-agent	1	.782
rest-agent	1	.782
search-agent	1	.879
sleep-agent	1	.782
wake_up-agent	1	.782

Table E2: *Performance of the Best Spatial and Best Graphical Model on the Selectional Task Using Indirect Word Pairs in Simulation 1*

	Agent		Patient	
Verb	Co-occurrence Graph	Co-occurrence Space	Co-occurrence Graph	Co-occurrence Space
butcher	.903	.701	.938	.861
catch	.903	.888	.938	.861
chase	.903	.888	.936	.861
cook	.903	.701	.938	.861
crack	.902	.698	.881	.573
drink	NA	NA	.665	.1
eat	NA	NA	.983	.701
gather	.903	.915	.990	.950
go_to	NA	NA	.928	.983
peel	.903	.522	.768	.417
reach	1	.969	.714	.562
shoot	.903	.701	.914	.564
stab	.903	.701	.869	.563
stretch	.939	.969	NA	NA
throw_(spear)_at	.903	.701	.926	.564
trap	.903	.701	.869	.563
yawn	.988	.997	NA	NA

Note. Performance is separated by each noun-verb pair tested. A cell is filled with NA if the verb does not have arguments with the corresponding thematic role, e.g. the verb *stretch* does not take any patient noun; or the nouns that the verb does not take as a direct argument never filled in the corresponding thematic role. For example, all the nouns that can be evaluated in the indirect condition for *eat* are the non-agent nouns like *apple* and *water*, which would never take the agent role, i.e. $apple_a$ and $water_a$ are not allowed due to the semantic constraints in Appendix A. In this case, *eat* only have nouns in agent role to evaluate in the direct condition, but not in the indirect condition.

Appendix F
Performance of Top Models by Encoding Type

Table F1 *Performance of the Best Models in 14 Conditions on the Selectional Task Using Direct Word Pairs.*

		Encoding Type					
		Unreduced Similarity (no SVD)			**Reduced Similarity (SVD)**		
Data Structure	**Co-occurrence**	**Distance**	**Cosine**	**Correlation**	**Distance**	**Cosine**	**Correlation**
Spatial	1	0.703	0.598	0.453	0.583	0.570	0.685
Graphical	1	0.703	0.598	0.453	0.583	0.570	0.685

Note. Performance is separated by Encoding Type.

Table F2 *Performance of the Best Models in 14 Conditions on the Selectional Task Using Indirect Word Pairs.*

		Encoding Type					
		Unreduced Similarity (no SVD)			**Reduced Similarity (SVD)**		
Data Structure	**Co-occurrence**	**Distance**	**Cosine**	**Correlation**	**Distance**	**Cosine**	**Correlation**
Spatial	0.758	0.723	0.722	0.713	0.715	0.695	0.702
Graphical	0.901	0.742	0.750	0.737	0.713	0.464	0.614