Vector-based Models of Semantic Composition

Jeff Mitchell and Mirella Lapata
School of Informatics, University of Edinburgh
10 Chrichton Street, Edinburgh EH10 9AB, UK
jeff.mitchell@ed.ac.uk, mlap@inf.ed.ac.uk

Abstract

Vector-based models of word meaning have become increasingly popular in cognitive science. The appeal of these models lies in their ability to represent meaning simply by using distributional information under the assumption that words occurring within similar contexts are semantically similar. Despite their widespread use, vector-based models are typically directed at representing words in isolation and methods for constructing representations for phrases or sentences have received little attention in the literature. This is in marked contrast to experimental evidence (e.g., in sentential priming) suggesting that semantic similarity is more complex than simply a relation between isolated words. This article proposes a framework for representing the meaning of word combinations in vector space. Central to our approach is vector composition which we operationalize in terms of additive and multiplicative functions. Under this framework, we introduce a wide range of composition models which we evaluate empirically on a phrase similarity task.

Introduction

The question of how semantic knowledge is acquired, organized, and ultimately used in language processing and understanding has been a topic of great debate in cognitive science. This is hardly surprising as the ability to retrieve and manipulate meaning influences many cognitive tasks that go far and beyond language processing. Examples include memory retrieval (Deese, 1959; Raaijmakers & Shiffrin, 1981), categorization (Estes, 1994; Nosofsky, 1984, 1986), problem solving (Holyoak & Koh, 1987; Ross, 1987, 1989a), reasoning (Rips, 1975; Heit & Rubinstein, 1994), and learning (Gentner, 1989; Ross, 1984).

Previous accounts of semantic representation fall under three broad families, namely semantic networks, feature-based models, and semantic spaces (for a fuller account of the different approaches and issue involved see e.g. Markman, 1998). Semantic networks (Collins & Quillian, 1969) represent concepts as nodes in a graph. Edges in the graph denote semantic relationships between concepts (e.g., DOG IS-A MAMMAL, DOG HAS TAIL) and word meaning is expressed by the number and type of connections to other words. In this framework, word similarity is a function of path length — semantically related words are expected to have shorter paths between them (e.g., *poodle* will be more similar to *dog* than *animal*). Semantic networks constitute a somewhat idealized representation that abstracts away from real word usage — they are traditionally

hand-coded by modelers who *a priori* decide which relationships are most relevant in representing meaning. More recent work (Steyvers & Tenenbaum, 2005) creates a semantic network from word association norms (Nelson, McEvoy, & Schreiber, 1999), however these can only represent a small fraction of the vocabulary of an adult speaker.

An alternative to semantic networks is the idea that word meaning can be described in terms of feature lists (Smith & Medin, 1981). Theories tend to differ with respect to their definition of features. In many cases these are created manually by the modeler (e.g., Hinton & Shallice, 1991). In other cases, the features are obtained by asking native speakers to generate attributes they consider important in describing the meaning of a word (e.g., Andrews, Vigliocco, & Vinson, 2009; McRae, de Sa, & Seidenberg, 1997). This allows the representation of each word by a distribution of numerical values over the feature set. Admittedly, norming studies have the potential of revealing which dimensions of meaning are psychologically salient. However, a number of difficulties arise when working with such data (Murphy & Medin, 1985; Sloman & Rips, 1998). For example, the number and types of attributes generated can vary substantially as a function of the amount of time devoted to each word. There are many degrees of freedom in the way that responses are coded and analyzed. And multiple subjects are required to create a representation for each word, which in practice limits elicitation studies to a small-size lexicon.

A third popular tradition of studying semantic representation has been driven by the assumption that word meaning can be learned from the linguistic environment. Words that are similar in meaning e.g., *boat* and *ship* tend to occur in contexts of similar words, such as *sail*, *sea*, *sailor*, and so on. Word meaning can be thus captured *quantitatively* in terms of simple co-occurrence statistics. *Semantic space* models thus represent meaning as a vector in a high-dimensional space, where each component corresponds to some contextual element in which the word is found. The contextual elements can be words themselves (Lund & Burgess, 1996), larger linguistic units such as paragraphs or documents (Landauer & Dumais, 1997), or even more complex linguistic representations such as *n*-grams (Jones & Mewhort, 2007) and the argument slots of predicates (Grefenstette, 1994; Lin, 1998; Padó & Lapata, 2007). The advantage of taking such a geometric approach is that the similarity of word meanings can be easily quantified by measuring their distance in the vector space, or the cosine of the angle between them. A simplified example of a two-dimensional semantic space is shown in Figure 1 (semantic spaces usually have hundreds of dimensions).

There are a number of well-known semantic space models in the literature. For example, the Hyperspace Analog to Language model (HAL, Lund & Burgess, 1996) represents each word by a vector where each element of the vector corresponds to a weighted co-occurrence value of that word with some other word. Latent Semantic Analysis (LSA, Landauer & Dumais, 1997) also derives a high-dimensional semantic space for words while using co-occurrence information between words and the passages they occur in. LSA constructs a word-document co-occurrence matrix from a large document collection. Matrix decomposition techniques are usually applied to reduce the dimensionality of the original matrix thereby rendering it more informative. The dimensionality reduction allows words with similar meaning to have similar vector representations, even if they never co-occurred in the same document.

Probabilistic topic models (Griffiths, Steyvers, & Tenenbaum, 2007; Blei, Ng, & Jordan, 2003) offer an alternative to semantic spaces based on the assumption that words observed in a corpus manifest some latent structure linked to topics. These models are similar in spirit to LSA, they also operate on large corpora and derive a reduced dimensionality description of words and documents. Crucially, words are not represented as points in a high-dimensional space but as a

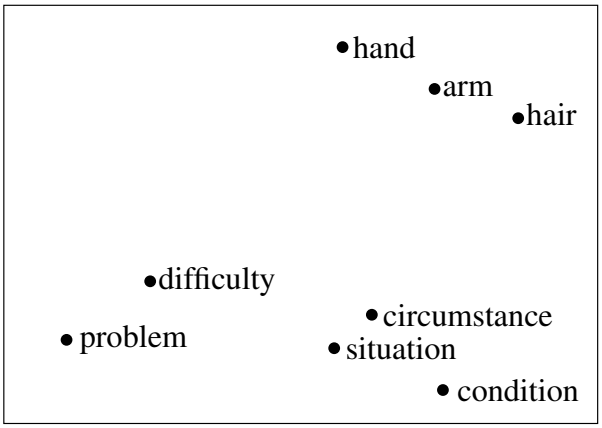


Figure 1. In a semantic space words are represented as points, and proximity indicates semantic association. Here, *circumstance, situation* and *condition* are similar to each other and different from *hand, arm* and *hair*.

probability distribution over a set of topics (corresponding to coarse-grained senses). Each topic is a probability distribution over words, and the content of the topic is reflected in the words to which it assigns high probability. Topic models are *generative*, they specify a probabilistic procedure by which documents can be generated. So, to make a new document one first chooses a distribution over topics. Then for each word in that document, one chooses a topic at random according to this distribution, and selects a word from that topic. Under this framework, the problem of meaning representation is expressed as one of statistical inference: given some data — words in a corpus — infer the latent structure from which it was generated.

Semantic space models (and the related topic models) have been successful at simulating a wide range of psycholinguistic phenomena including semantic priming (Lund & Burgess, 1996; Landauer & Dumais, 1997; Griffiths et al., 2007), discourse comprehension (Landauer & Dumais, 1997; Foltz, Kintsch, & Landauer, 1998), word categorization (Laham, 2000), judgments of essay quality (Landauer, Laham, Rehder, & Schreiner, 1997a), synonymy tests (Landauer & Dumais, 1997; Griffiths et al., 2007) such as those included in the Test of English as Foreign Language (TOEFL), reading times (McDonald, 2000; Griffiths et al., 2007), and judgments of semantic similarity (McDonald, 2000) and association (Denhire & Lemaire, 2004; Griffiths et al., 2007).

Despite their widespread use, these models are typically directed at representing words in isolation and methods for constructing representations for phrases or sentences have received little attention in the literature. However, it is well-known that linguistic structures are *compositional* (simpler elements are combined to form more complex ones). For example, morphemes are combined into words, words into phrases, and phrases into sentences. It is also reasonable to assume that the meaning of sentences is composed of the meanings of individual words or phrases. Much experimental evidence also suggests that semantic similarity is more complex than simply a relation between isolated words. For example, Duffy, Henderson, and Morris (1989) showed that priming of sentence terminal words was dependent not simply on individual preceding words but on their combination, and Morris (1994) later demonstrated that this priming also showed dependencies on

the syntactic relations in the preceding context. Additional evidence comes from experiments where target words in sentences are compared to target words in lists or scrambled sentences. Changes in the temporal order of words in a sentence decrease the strength of the related priming effect (Foss, 1982; Masson, 1986; O'Seaghdha, 1989; Simpson, Peterson, Casteel, & Brugges, 1989). For example, Simpson et al. (1989) found relatedness priming effects for words embedded in grammatical sentences (*The auto accident drew a large crowd of people*) but not for words in scrambled stimuli (*Accident of large the drew auto crowd a people*). These findings highlight the role of syntactic structure in modulating priming behavior. They also suggest that models of semantic similarity should ideally handle the combination of semantic content in a syntactically aware manner.

Composition operations can be naturally accounted for within logic-based semantic frameworks (Montague, 1974). Frege's principle of compositionality states that the meaning of a complete sentence must be explained in terms of the meanings of its subsentential parts, including those of its singular terms. In other words, each syntactic operation of a formal language should have a corresponding semantic operation. Problematically, representations in terms of logical formulas are not well suited to modeling similarity quantitatively (as they are based on discrete symbols). On the other hand, semantic space models can naturally measure similarity but are not compositional. In fact, the commonest method for combining the vectors is to average them. While vector averaging has been effective in some applications such as essay grading (Landauer & Dumais, 1997) and coherence assessment (Foltz et al., 1998), it is unfortunately insensitive to word order, and more generally syntactic structure, giving the same representation to any constructions that happen to share the same vocabulary. This is illustrated in the example below taken from Landauer, Laham, Rehder, and Schreiner (1997b). Sentences (1-a) and (1-b) contain exactly the same set of words but their meaning is entirely different.

(1) a. It was not the sales manager who hit the bottle that day, but the office worker with the serious drinking problem.
    b. That day the office manager, who was drinking, hit the problem sales worker with the bottle, but it was not serious.

The relative paucity of compositional models in the semantic space literature is in marked contrast with work in the connectionist tradition where much effort has been devoted to problem of combining or *binding* high-dimensional representations. The construction of higher-level structures from low-level ones is fundamental not only to language but many aspects of human cognition such as analogy retrieval and processing (Plate, 2000; Eliasmith & Thagard, 2001), memory (Kanerva, 1988), and problem solving (Ross, 1989b). Indeed, the issue of how to represent compositional structure in neural networks has been a matter of great controversy (Fodor & Pylyshyn, 1988). While neural networks can readily represent single distinct objects, in the case of multiple objects there are fundamental difficulties in keeping track of which features are bound to which objects. For the hierarchical structure of natural language this binding problem becomes particularly acute. For example, simplistic approaches to handling sentences such as *John loves Mary* and *Mary loves John* typically fail to make valid representations in one of two ways. Either there is a failure to distinguish between these two structures, because the network fails to keep track of the fact that *John* is subject in one and object in the other, or there is a failure to recognize that both structures involve the same participants, because *John* as a subject has a distinct representation from *John* as an object. The literature is littered with solutions to the binding problem (see the following section for a detailed overview). These include tensor products (Smolensky, 1990), recursive distributed

representations (RAAMS, Pollack, 1990), spatter codes (Kanerva, 1988), holographic reduced representations (Plate, 1995), and convolution (Metcalfe, 1990).

In this article, we attempt to bridge the gap in the literature by developing models of semantic composition that can represent the meaning of word combinations as opposed to individual words. Our models are narrower in scope compared to those developed in earlier connectionist work. Our vectors represent words, they are high-dimensional but relatively structured, every component corresponds to a predefined context in which the words are found. We take it as a defining property of the vectors we consider that the values of their components are derived from event frequencies such as the number of times a given word appears in a given context.[1] Having this in mind, we present a general framework for vector-based composition which allows us to consider different classes of models. Specifically, we formulate composition as a function of two vectors and introduce models based on addition and multiplication. The similarity between two complex expressions can be thus naturally expressed using a geometric measure such as cosine or Euclidean distance. We also investigate how the choice of the underlying semantic representation interacts with the choice of composition function by comparing a spatial model that represents words as vectors in a high-dimensional space against a probabilistic model that represents words as topic distributions. We assess the performance of these models directly on a similarity task. We elicit similarity ratings for pairs of adjective-noun, noun-noun and verb-object constructions and examine the strength of the relationship between similarity ratings and the predictions of our models.

In the remainder, we review previous research on semantic composition and vector binding models. Next, we describe our modeling framework, present our elicitation experiments, and discuss our results.

Composition

Compositionality allows languages to construct complex meanings from combinations of simpler elements. This property is often captured in the following principle: the meaning of a whole is a function of the meaning of the parts (Partee, 1995, p. 313). Therefore, whatever approach we take to modeling semantics, representing the meanings of complex structures will involve modeling the way in which meanings combine. Let us express the composition of two constituents, $\mathbf{u}$ and $\mathbf{v}$, in terms of a function acting on those constituents:

$$\mathbf{p} = f(\mathbf{u}, \mathbf{v}) \tag{1}$$

Partee (1995, p. 313) suggests a further refinement of the above principle taking the role of syntax into account: the meaning of a whole is a function of the meaning of the parts and of the way they are syntactically combined. We thus modify the composition function in (1) to account for the fact that there is a syntactic relation R between constituents $\mathbf{u}$ and $\mathbf{v}$:

$$\mathbf{p} = f(\mathbf{u}, \mathbf{v}, R) \tag{2}$$

Unfortunately, even this formulation may not be fully adequate. Lakoff (1977, p. 239), for example, suggests that the meaning of the whole is greater than the meaning of the parts. The

[1] For example, vectors offer a convenient representation for encoding features in machine learning, however the values of these vectors are not always derived from event frequencies. Graphs are also often represented by an adjacency matrix, a matrix with rows and columns labeled by graph vertices v, with a 1 or 0 in position (v_i, v_j) according to whether v_i and v_j are adjacent or not. This does not imply that an adjacency matrix is a vector-based model, as the values of the elements in the matrix do not correspond to event frequencies.

implication here is that language users are bringing more to the problem of constructing complex meanings than simply the meaning of the parts and their syntactic relations. This additional information includes both knowledge about the language itself and also knowledge about the real world. Thus, full understanding of the compositional process involves an account of how novel interpretations are integrated with existing knowledge. Again, the composition function needs to be augmented to include an additional argument, K, representing any knowledge utilized by the compositional process:

$$\mathbf{p} = f(\mathbf{u}, \mathbf{v}, R, K) \tag{3}$$

The difficulty of defining compositionality is highlighted by Frege (1884, p. x) himself who cautions never to ask for the meaning of a word in isolation but only in the context of a statement. In other words, it seems that the meaning of the whole is constructed from its parts, and the meaning of the parts is derived from the whole. Moreover, compositionality is a matter of degree rather than a binary notion. Linguistic structures range from fully compositional (e.g., *black hair*), to partly compositional syntactically fixed expressions, (e.g., *take advantage*), in which the constituents can still be assigned separate meanings, and non-compositional idioms (e.g., *kick the bucket*) or multiword expressions (e.g., *by and large*), whose meaning cannot be distributed across their constituents (Nunberg, Sag, & Wasow, 1994).

Despite the foundational nature of compositionality to language, there are significant obstacles to understanding what exactly it is and how it operates. Most significantly, there is the fundamental difficulty of specifying what sort of "function of the meanings of the parts" is involved in semantic composition (Partee, 2004, p. 153). Fodor and Pylyshyn (1988) attempt to characterize this function by appealing to the notion of *systematicity*. They argue that the ability to understand some sentences is intrinsically connected to the ability to understand certain others. For example, no-one who understands *John loves Mary* fails to understand *Mary loves John*. Therefore, the semantic content of a sentence is systematically related to the content of its constituents and the ability to recombine these according to a set of rules. In other words, if one understands some sentence and the rules that govern its construction, one can understand a different sentence made up of the same elements according to the same set of rules. In a related proposal, Holyoak and Hummel (2000) claim that in combining parts to form a whole, the parts remain independent and maintain their identities. This entails that *John* has the same independent meaning in both *John loves the girl* and *The boy hates John*.

Aside from the philosophical difficulties of precisely determining what systematicity means in practice (Pullum & Scholz, 2007; Spenader & Blutner, 2007; Doumas & Hummel, 2005), it is worth noting that semantic transparency, the idea that words have meanings which remain unaffected by their context, contradicts Frege's (1884) claim that words only have definite meanings in context. Consider for example the adjective *good* whose meaning is modified by the context in which it occurs. The sentences *John is a good neighbor* and *John is a lawyer* do not imply *John is a good lawyer*. In fact, we might expect that some of the attributes of a good lawyer are incompatible with being a good neighbor, such as nit-picking over details, or not giving an inch unless required by law. More generally, the claims of Fodor and Pylyshyn (1988) and Holyoak and Hummel (2000) arise from a preconception of cognition as being essentially symbolic in character. While it is true that the concatenation of any two symbols (e.g., g and l), will compose into an expression (e.g., gl), within which both symbols maintain their identities, we cannot always assume that the the meaning of a phrase is derived by simply concatenating the meaning of its constituents. Although the phrase *good lawyer* is constructed by concatenating the symbols *good* and *lawyer*, the meaning of *good*

will vary depending on the nouns it modifies.

Interestingly, Pinker (1994, p. 84) discusses the types of functions that are *not* involved in semantic composition while comparing languages, which he describes as *discrete combinatorial systems*, against blending systems. He argues that languages construct an unlimited number of completely distinct combinations with an infinite range of properties. This is made possible by creating novel, complex meanings which go beyond those of the individual elements. In contrast, for a blending system the properties of the combination lie between the properties of its elements, which are lost in the average or mixture. To give a concrete example, a *brown cow* does not identify a concept intermediate between *brown* and *cow* (Kako, 1999, p. 2). Thus, composition based on averaging or blending would produce greater generality rather than greater specificity.

Logic-based View Within symbolic logic, compositionality is accounted for elegantly by assuming a tight correspondence between syntactic expressions and semantic form (Montague, 1974; Blackburn & Bos, 2005). In this tradition, the meaning of a phrase or sentence is its truth-conditions which are expressed in terms of truth relative to a model.[2] In classical Montague grammar, for each syntactic category there is a uniform semantic type (e.g., sentences express propositions, nouns and adjectives express properties of entities, and verbs express properties of events). Most lexical meanings are left unanalyzed and treated as primitive. In this framework, the proper noun *John* is represented by the logical symbol *JOHN* denoting a specific entity, whereas a verb like *wrote*, is represented by a function from entities to propositions, expressed in lambda calculus as $\lambda x.WROTE(x)$. Applying this function to the entity *JOHN* yields the logical formula $WROTE(JOHN)$ as a representation of the sentence *John wrote*. It is worth noting that the entity and predicate within this formula are represented symbolically, and that the connection between a symbol and its meaning is an arbitrary matter of convention.

On one hand, the symbolic nature of logical representations is advantageous as it allows composition to be carried out syntactically. The laws of deductive logic in particular can be defined as syntactic processes which act irrespective of the meanings of the symbols involved. On the other hand, abstracting away from the actual meanings may not be fully adequate for modeling semantic composition. For example, adjective-noun phrases are represented in terms of predicate conjunction, e.g., *male lawyer* corresponds to $MALE(x) \wedge LAWYER(x)$. This approach cannot, however, handle the context sensitive adjectives discussed above. *John is a good lawyer* is not equivalent to the conjunction of *John is good* and *John is a lawyer*. More generally, modeling semantic composition means modeling the way in which meanings combine, and this requires that words have representations which are richer than single, arbitrary symbols.

Connectionism Connectionist models of cognition (see among others Elman et al., 1996; Rumelhart, McClelland, & the PDP Research Group, 1986) can be seen as a response to the limitations of traditional symbolic models. The key premise here is that knowledge is represented not as discrete symbols that enter into symbolic expressions, but as patterns of activation distributed over many processing elements. These representations are distributed in the sense that any single concept is

[2]The structure common to all of the models in which a given language is interpreted reflects certain basic presuppositions about the "structure of the world" that are implicit in the language. In predicate logic, a model consists of the set of truth-values $\{0,1\}$, a domain D which is some set of entities, and some n-ary relations on this set. The model also consists of an interpretation function which assigns semantic values to all constants.

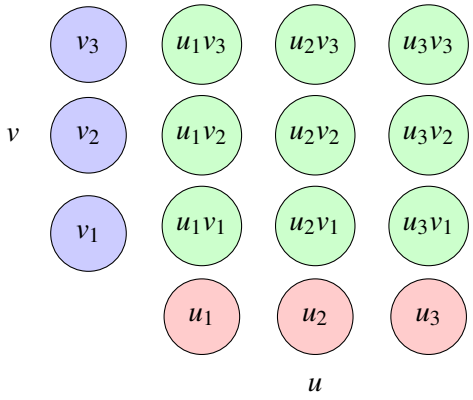


Figure 2. The tensor product of two three-dimensional vectors u and v.

represented as a pattern, i.e., vector, of activation over many elements (nodes or units) that are typically assumed to correspond to neurons or small collections of neurons.

Much effort in the literature has been invested in enhancing the representational capabilities of connectionist models with the means to combine a finite number of symbols into a much larger, possibly infinite, number of specific structures. The key property of symbolic representations that connectionist models attempt to emulate is their ability to bind one representation to another. The fundamental operation underlying binding in symbolic systems is the concatenation of symbols according to certain syntactic processes. And crucially the results of this operation can be broken down into their original constituents. Thus, connectionists have sought ways of constructing complex structures by binding one distributed representation to another in a manner that is reversible.

Smolensky (1990), for example, proposed the use of tensor products as a means of binding one vector to another to produce structured representations. The tensor product $\mathbf{u} \otimes \mathbf{v}$ is a matrix whose components are all the possible products $u_i v_j$ of the components of vectors $\mathbf{u}$ and $\mathbf{v}$. Figure (1) illustrates the tensor product for two three-dimensional vectors $(u_1, u_2, u_3) \otimes (v_1, v_2, v_3)$. A major difficulty with tensor products is their dimensionality which grows exponentially in size as more constituents are composed (precisely, the tensor product has dimensionality $m \times n$).

To overcome this problem, other techniques have been proposed in which the binding of two vectors results in a vector which has the same dimensionality as its components. Holographic reduced representations (Plate, 1991) are one implementation of this idea where the tensor product is projected onto the space of the original vectors, thus avoiding any dimensionality increase. The projection is defined in terms of *circular convolution* a mathematical function that compresses the tensor product of two vectors. The compression is achieved by summing along the transdiagonal elements of the tensor product. Noisy versions of the original vectors can be recovered by means of *circular correlation* which is the approximate inverse of circular convolution. The success of circular correlation crucially depends on the components of the n-dimensional vectors $\mathbf{u}$ and $\mathbf{v}$ being real numbers and randomly distributed with mean 0 and variance $\frac{1}{n}$. Binary spatter codes (Kanerva, 1988, 2009) are a particularly simple form of holographic reduced representation. Typically, these vectors are random bit strings or binary N-vectors (e.g., $N = 10,000$). Compositional representations are synthesized from parts or chunks. Chunks are combined by binding which is the same as

taking the exclusive or (XOR) of two vectors. Here, only the transdiagonal elements of the tensor product of two vectors are kept, and the rest are discarded.

From a computational perspective, both spatter codes and holographic reduced representations can be implemented efficiently[3] and the dimensionality of the resulting vector does not change. The downside is that operations like circular convolution are a form of lossy compression that introduces noise into the representation. To retrieve the original vectors from their bindings a *clean-up memory* process is usually employed where the noisy vector is compared to all component vectors in order to find the closest one.

Tensors and their relatives can indeed represent relations (e.g., *love(x,y)*) and role-filler bindings (e.g., in *loves(John, Mary)* the *lover* role is bound to *John* and the *beloved* role is bound to *Mary*) in a distributed fashion. However, Holyoak and Hummel (2000) claim that this form of binding violates role-filler independence. In a truly compositional system, complex structures gain meaning from the simpler parts from which they are formed *and* the simpler components remain independent, i.e., preserve their meaning (Doumas & Hummel, 2005; Doumas, Hummel, & Sandhofer, 2008). Doumas and Hummel (2005) propose a model of role-filler binding based on synchrony of neural firing. Vectors representing relational roles fire in synchrony with vectors representing their fillers and out of synchrony with other role-filler bindings. These ideas are best captured in LISA, a neural network that implements symbolic structures in terms of distributed representations. Crucially, words and relations are represented by features (e.g., *human*, *adult*, *male*) which albeit more informative than binary vectors, raise issue regarding their provenance and the scalability of the models based on them (see the discussion in the Introduction).

Semantic Spaces The idea of representing word meaning in a geometrical space dates back to Osgood, Suci, and Tannenbaum (1957), who used elicited similarity judgments to construct semantic spaces. Subjects rated concepts on a series of scales whose endpoints represented polar opposites (e.g., *happy–sad*); these ratings were further processed with factor analysis, a dimensionality reduction technique, to uncover latent semantic structure. In this study, meaning representations were derived *directly* from psychological data, thereby allowing the analysis of differences across subjects. Unfortunately, multiple subject ratings are required to create a representation for each word, which in practice limits the semantic space to a small number of words.

Building on this work and the well-known vector space model in information retrieval (Salton, Wong, & Yang, 1975; Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990), more recent semantic space models, such as LSA (Landauer & Dumais, 1997) and HAL lundburgess1996, overcome this limitation by constructing semantic representations *indirectly* from real language corpora. A variety of such models have been proposed and evaluated in the literature. Despite their differences, they are all based on the same premise: words occurring within similar contexts are semantically similar (Harris, 1968). Semantic space models extract from a corpus a set of counts representing the occurrences of a target word t in the specific context c of choice and then map these counts into the components of a vector in some space. For example, Bullinaria and Levy (2007) consider a range of component types, the simplest being to transform the raw frequencies into conditional probabilities, $p(c_i|t)$. They also consider components based on functions of these probabilities, such as the ratio of the conditional probability of the context to its overall probability, or the pointwise mutual information between context and target. An issue here concerns the

[3] Binary spatter codes have a runtime complexity of $O(N)$ for a vector of length N and holographic reduced representations can be implemented using the Fast Fourier Transform, which is $O(N \log N)$.

number of components the vectors should have, or which contexts should be used in constructing the vectors. Often, the most frequent contexts are used, as rarer contexts yield unreliable counts. Dimensionality reduction techniques can be also used to project high dimensional vectors onto a lower dimensional space (Landauer & Dumais, 1997; Hofmann, 2001; Blei et al., 2003).

Semantic space models resemble the representations used in the connectionist literature. Words are represented as vectors and their meaning is distributed across many dimensions. Crucially, the vector components are neither binary nor randomly distributed. They correspond to co-occurrence counts and it is assumed that differences in meaning arise from differences in the distribution of these counts across contexts. That is not to say that high-dimensional randomly distributed representations are incompatible with semantic spaces. Kanerva, Kristoferson, and Holst (2000) propose the use of random indexing as an alternative to the computationally costly singular value decomposition employed in LSA. The procedure also builds a word-document co-occurrence matrix, except that each document no longer has its own column. Instead, it is assigned a small number of columns at random (the document's random index). So, each time a word occurs in the document, the document's random index vector is added to the row corresponding to that word.

Random vectors have also been employed in an attempt to address a commonly raised criticism against semantic space models, namely that they are inherently agnostic to the linguistic structure of the contexts in which a target word occurs. In other words, most of these models treat these contexts as a structureless bag-of-words. Jones and Mewhort (2007) propose a model that makes use of the linear order of words in a context. Their model represents words by high-dimensional holographic vectors. Each word is is assigned a random[4] *environmental* vector. Contextual information is stored in a lexical vector which is computed with the aid of the environmental vectors. Specifically, a word's lexical vector is the superposition of the environmental vectors corresponding to its co-occurring words in a sentence. Order information is the sum of all n-grams that include the target word. The n-grams are encoded with the aid of a place-holder environmental vector Φ and circular convolution (Plate, 1995). The order vector is finally added to the lexical vector in order to jointly represent structural and contextual information. Despite the fact that these vectors contain information about multi-word structures in the contexts of target words, they are, nonetheless, still fundamentally representations of individual isolated target words. Circular convolution is only used to bind environmental vectors, which being random contain no semantic information. To make a useful semantic representation of a target word, the vectors representing its contexts are summed over, producing a vector which is no longer random and for which circular convolution is no longer optimal.

Sahlgren, Host, and Kanerva (2008) provide and alternative to convolution by showing that order information can also be captured by permuting the vector coordinates. Other models implement more sophisticated versions of context that go beyond the bag-of-words model, without however resorting to random vectors. For example, by defining context in terms of syntactic dependencies (Grefenstette, 1994; Lin, 1998; Padó & Lapata, 2007) or by taking into account relational information about how roles and fillers combine to create specific factual knowledge (Dennis, 2007).

So far the discussion has centered on the creation of semantic representations for individual words. As mentioned earlier, the composition of vector-based semantic representations has received relatively little attention. An alternative is not to compose at all but rather create semantic representations for phrases in addition to words. If a phrase is frequent enough, then it can be treated as a

[4] Vector components are sampled at random from a Gaussian distribution with $\mu = 0$ and $\sigma = \frac{1}{\sqrt{D}}$ where $D = 2,048$.

single target unit, and its occurrence across a range of contexts can be constructed in the same manner as described above. Baldwin, Bannard, Tanaka, and Widdows (2003) apply this method to model the decomposability of multi-word expressions such as noun compounds and phrasal verbs. Taking a similar approach, Bannard, Baldwin, and Lascarides (2003) develop a vector space model for representing the meaning of verb-particle constructions. In the limit, such an approach is unlikely to work as semantic representations for constructions that go beyond two-words will be extremely sparse.

Vector addition or averaging (which are equivalent under the cosine measure) is the most common form of vector combination (Landauer & Dumais, 1997; Foltz et al., 1998) However, vector addition is not a suitable model of composition for at least two reasons. Firstly, it is insensitive to syntax and word order. Because vector addition is commutative, it assigns the same representation to any sentence containing the same constituents irrespective of their syntactic relations. It is therefore a bag-of-words model of composition. In contrast, there is ample empirical evidence that syntactic relations across and within sentences are crucial for sentence and discourse processing (Neville, Nichol, Barss, Forster, & Garrett, 1991; West & Stanovich, 1986). Secondly, addition simply blends together the content of all words involved to produce something in between them all. Ideally, we would like a model of semantic composition that generates novel meanings by selecting and modifying particular aspects of the constituents participating in the composition. Kintsch (2001) attempts to achieve this in his predication algorithm by modeling how the meaning of a predicate (e.g., *run*) varies depending on the arguments it operates upon (e.g, *the horse ran* vs. *the color ran*). The idea is to add not only the vectors representing the predicate and its argument but also the neighbors associated with both of them. The neighbors, Kintsch argues, can strengthen features of the predicate that are appropriate for the argument of the predication.

Tensor products have been recently proposed as an alternative to vector addition. (Aerts & Czachor, 2004; Clark & Pulman, 2007; Widdows, 2008). However, as illustrated in Figure (1), these representations grow exponentially as more vectors are combined. This fact undermines not only their tractability in an artificial computational setting but also their plausibility as models of human concept combination. Interestingly, Clark, Coecke, and Sadrzadeh (2008) try to construct a tensor product based model of vector composition which makes an explicit connection to models of linguistic composition. In particular, they show how vector-based semantics can be unified with a compositional theory of grammatical types. Central to their approach is the association of each grammatical type with a particular rank of tensor. So, for example, if we take nouns as being associated with simple vectors, then an adjective as a noun modifier would be associated with a matrix, i.e. a vector transformation. Clark et al. (2008) do not suggest concrete methods for constructing or estimating the various tensors involved in their model. Instead, they are more interested in its formal properties and do not report any empirical tests of this approach.

Unfortunately, comparisons across vector composition models have been few and far between. The merits of different approaches are illustrated with special purpose examples and large scale evaluations are uniformly absent. For instance, Kintsch (2001) demonstrates how his own composition algorithm works intuitively on a few hand selected examples but does not provide a comprehensive test set (see Frank, Koppen, Noordman, & Vonk, 2007 for a criticism of Kintsch's 2001 evaluation standards). In a similar vein, Widdows (2008) explores the potential of vector product operations for modeling compositional phenomena in natural language, again on a small number of hand picked examples.

Our work goes beyond these isolated proposals; we present a framework for vector com-

	music	solution	economy	craft	reasonable
practical	0	6	2	10	4
difficulty	1	8	4	4	0

Figure 3. A hypothetical semantic space for *practical* and *difficulty*.

position which allows us to explore a range of potential composition functions, their properties, and relations. Under this framework, we reconceptualize existing composition models as well as introduce novel ones. Our experiments make use of conventional semantic vectors built from co-occurrence data. However, our compositional models are not tied to a specific representation and could be used with the holographic vectors proposed in Jones and Mewhort (2007) or with random indexing, however we leave this to future work. Within the general framework of co-occurrence-based models we investigate how the choice of semantic representation interacts with the choice of composition model. Specifically, we compare a spatial model that represents words as vectors in a high-dimensional space against a probabilistic model (akin to LSA) that represents words as topic distributions. We evaluate these models empirically on a phrase similarity task, using a rigorous evaluation methodology.

Composition Models

Our aim is to construct vector representations for phrases and sentences. We assume that constituents are represented by vectors which subsequently combine in some way to produce a new vector. It is worth emphasizing that the problem of combining semantic vectors to make a representation of a multi-word phrase, is different to the problem of how to incorporate information *about* multi-word contexts into a distributional representation for a single target word. Whereas Jones and Mewhort (2007) test this ability to memorize the linear structure of contexts in terms of predicting a target word correctly given a context, our composition models will be evaluated in terms of their ability to model semantic properties of simple phrases.

In this study we focus on small phrases, consisting of a head and a modifier or complement, which form the building blocks of larger units. If we cannot model the composition of basic phrases, there is little hope that we can construct compositional representations for sentences or even documents (we return to this issue in our Discussion section). So, given a phrase such as *practical difficulty* and the vectors $\mathbf{u}$ and $\mathbf{v}$ representing the constituents *practical* and *difficulty*, respectively, we wish to produce a representation $\mathbf{p}$ of the whole phrase. Hypothetical vectors for these constituents are illustrated in Figure 3. This simplified semantic space[5] will serve to illustrate examples of the composition functions we consider in this paper.

In our earlier discussion, we defined $\mathbf{p}$, the composition of vectors $\mathbf{u}$ and $\mathbf{v}$, representing a pair of words which stand in some syntactic relation R, given some background knowledge K as:

$$\mathbf{p} = f(\mathbf{u}, \mathbf{v}, R, K) \tag{4}$$

The expression above defines a wide class of composition functions. To derive specific models from this general framework requires the identification of appropriate constraints that narrow the space of

[5]The space has only five dimensions; the matrix cells denote the co-occurrence of the words *practical* and *difficulty* with *music*, *solution*, and so on. We also experiment with an alternative semantic representation denoting the distribution of words over topics. We refer the reader to our modeling experiments for details.

functions being considered. To begin with, we will ignore K so as to explore what can be achieved in the absence of any background or world knowledge. While background knowledge undoubtedly contributes to the compositional process, and resources like WordNet (Fellbaum, 1998) may be used to provide this information, from a methodological perspective it is preferable to understand the fundamental processes of how representations are composed before trying to understand the interaction between existing representations and those under construction. As far as the syntactic relation R is concerned, we can proceed by investigating one such relation at a time, thus removing any explicit dependence on R, but allowing the possibility that we identify distinct composition functions for distinct syntactic relations.

Another particularly useful constraint is to assume that $\mathbf{p}$ lies in the same space as $\mathbf{u}$ and $\mathbf{v}$. This essentially means that all syntactic types have the same dimensionality. The simplification may be too restrictive as it assumes that verbs, nouns and adjectives are substantially similar enough to be represented in the same space. Clark et al. (2008) suggest a scheme in which the structure of a representation depends on its syntactic type, such that, for example, if nouns are represented by plain vectors then adjectives, as modifiers of nouns, are represented by matrices. More generally, we may question whether representations in a fixed space are flexible enough to cover the full expressivity of language. Intuitively, sentences are more complex than individual phrases and this should be reflected in the representation of their meaning. In restricting all representations within a space of fixed dimensions, we are implicitly imposing a limit on the complexity of structures which can be fully represented. Nevertheless, the restriction renders the composition problem computationally feasible. We can use a single method for constructing representations, rather than different methods for different syntactic types. In particular, constructing a vector of n elements is easier than constructing a matrix of n^2 elements. Moreover, our composition and similarity functions only have to apply to a single space, rather than a set of spaces of varying dimensions.

Given these simplifying assumptions, we can now begin to identify specific mathematical types of functions. For example, if we wish to work with linear composition functions, there are two ways to achieve this. We may assume that $\mathbf{p}$ is a linear function of the Cartesian product of $\mathbf{u}$ and $\mathbf{v}$, giving an additive class of composition functions:

$$\mathbf{p} = \mathbf{A}\mathbf{u} + \mathbf{B}\mathbf{v} \tag{5}$$

where $\mathbf{A}$ and $\mathbf{B}$ are matrices which determine the contributions made by $\mathbf{u}$ and $\mathbf{v}$ to $\mathbf{p}$.

Or, we can assume that $\mathbf{p}$ is a linear function of the tensor product of $\mathbf{u}$ and $\mathbf{v}$, giving a multiplicative class of composition functions:

$$\mathbf{p} = \mathbf{C}\mathbf{u}\mathbf{v} \tag{6}$$

where $\mathbf{C}$ is a tensor of rank 3, which projects the tensor product of $\mathbf{u}$ and $\mathbf{v}$ onto the space of $\mathbf{p}$. (For readers unfamiliar with vector and tensor algebra we provide greater detail in Appendix C).

Linearity is very often a useful assumption because it constrains the problem considerably. However, this usually means that the solution arrived at is an approximation to some other, non-linear, structure. Going beyond the linear class of multiplicative functions, we will also consider some functions which are quadratic in $\mathbf{u}$, having the general form:

$$\mathbf{p} = \mathbf{D}\mathbf{u}\mathbf{u}\mathbf{v} \tag{7}$$

where $\mathbf{D}$ is now a rank 4 tensor which projects the product $\mathbf{uuv}$ onto the space of $\mathbf{p}$.

Within the additive model class (equation (5)), the simplest composition function is vector addition:

$$\mathbf{p} = \mathbf{u} + \mathbf{v} \tag{8}$$

So, according to equation (8), the addition of the two vectors representing *practical* and *difficulty* (see Figure 3) would be $\mathbf{practical} + \mathbf{difficulty} = [1\ \ 14\ \ 6\ \ 14\ \ 4]$. This model assumes that composition is a symmetric function of the constituents; in other words, the order of constituents essentially makes no difference. While this might be reasonable for certain structures, a list perhaps, a model of composition based on syntactic structure requires some way of differentiating the contributions of each constituent.

Kintsch (2001) attempts to model the composition of a predicate with its argument in a manner that distinguishes the role of these constituents, making use of the lexicon of semantic representations to identify the features of each constituent relevant to their combination. Specifically, he represents the composition in terms of a sum of predicate, argument and a number of neighbors of the predicate.

$$\mathbf{p} = \mathbf{u} + \mathbf{v} + \sum_i \mathbf{n}_i \tag{9}$$

Considerable latitude is allowed in selecting the appropriate neighbors. Kintsch (2001) considers only the m most similar neighbors to the predicate, from which he subsequently selects k, those most similar to its argument. So, if in the composition of *practical* with *difficulty*, the chosen neighbor is *problem*, with $\mathbf{problem} = [2\ \ 15\ \ 7\ \ 9\ \ 1]$, then this produces the representation $\mathbf{practical} + \mathbf{difficulty} + \mathbf{problem} = [3\ \ 29\ \ 13\ \ 23\ \ 5]$.

This composition model draws inspiration from the construction-integration model (Kintsch, 1988), which was originally based on symbolic representations, and introduces a dependence on syntax by distinguishing the predicate from its argument. In this process the selection of relevant neighbors for the predicate plays a role similar to the integration of a representation with existing background knowledge in the original construction-integration model. Here, background knowledge takes the form of the lexicon from which the neighbors drawn.

A simpler approach to introducing dependence on the syntactic relation, R, is to weight the constituents differentially in the summation.

$$\mathbf{p} = \alpha\mathbf{v} + \beta\mathbf{u} \tag{10}$$

This makes the composition function asymmetric in $\mathbf{u}$ and $\mathbf{v}$ allowing their distinct syntactic roles to be recognized. For instance, we could give greater emphasis to heads than other constituents. As an example, if we set α to 0.4 and β to 0.6, then $0.4 \cdot \mathbf{practical} = [0\ \ 2.4\ \ 0.8\ \ 4\ \ 1.6]$ and $0.6 \cdot \mathbf{difficulty} = [0.6\ \ 4.8\ \ 2.4\ \ 2.4\ \ 0]$, and *practical difficulty* is represented by their sum $0.4 \cdot \mathbf{practical} + 0.6 \cdot \mathbf{difficulty} = [0.6\ \ 5.6\ \ 3.2\ \ 6.4\ \ 1.6]$.

An extreme form of this differential in the contribution of constituents is where one of the vectors, say $\mathbf{u}$, contributes nothing at all to the combination:[6]

$$\mathbf{p} = \mathbf{v} \tag{11}$$

In this case *practical difficulty* would be simply represented by $\mathbf{difficulty} = [1\ \ 8\ \ 4\ \ 4\ \ 0]$. Admittedly the model in (11) is impoverished and rather simplistic, however it can serve as a simple baseline against which to compare more sophisticated models.

[6] The model in (11) is equivalent to setting $\beta = 0$.

So far we have considered solely additive composition models. These models blend together the content of the constituents being composed. The contribution of **u** in equation (8) is unaffected by its relation to **v**. It might be preferable to scale each component of **u** with its relevance to **v**, namely to pick out the content of each representation that is relevant to their combination. This can be achieved by using a multiplicative function instead:

$$\mathbf{p} = \mathbf{u} \odot \mathbf{v} \tag{12}$$

where the symbol $\odot$ represents multiplication of the corresponding components:

$$p_i = u_i \cdot v_i \tag{13}$$

For this model, our example vectors would combine to give: $\mathbf{practical} \odot \mathbf{difficulty} = [0 \;\; 48 \;\; 8 \;\; 40 \;\; 0]$.

Note that the multiplicative function in (12) is still a symmetric function and thus does not take word order or syntax into account. However, equation (12) is a particular instance of the more general class of multiplicative functions (equation (6)), which allows the specification of asymmetric syntax-sensitive functions. For example, the tensor product is an instance of this class with **C** being the identity matrix.

$$\mathbf{p} = \mathbf{u} \otimes \mathbf{v} \tag{14}$$

Where the symbol $\otimes$ stands for the operation of taking all pairwise products of the components of **u** and **v**:

$$p_{i,j} = u_i \cdot v_j \tag{15}$$

So, the tensor product representation of *practical difficulty* is:

$$\mathbf{practical} \otimes \mathbf{difficulty} = \begin{matrix} 0 & 0 & 0 & 0 & 0 \\ 6 & 48 & 24 & 24 & 0 \\ 2 & 16 & 8 & 8 & 0 \\ 10 & 80 & 40 & 40 & 0 \\ 4 & 32 & 16 & 16 & 0 \end{matrix} \tag{16}$$

Circular convolution is also a member of this class:

$$\mathbf{p} = \mathbf{u} \circledast \mathbf{v} \tag{17}$$

where the symbol $\circledast$ stands for a compression of the tensor product based on summing along its transdiagonal elements:

$$p_i = \sum_j u_j \cdot v_{(i-j) mod_n} \tag{18}$$

Circular convolution compresses the matrix in (16) into the vector $\mathbf{practical} \circledast \mathbf{difficulty} = [116 \;\; 50 \;\; 66 \;\; 62 \;\; 80]$.

One reason for choosing such multiplicative functions is that the magnitudes of **u** and **v** can only affect the magnitude of **p**, not its direction. In contrast, in additive models, the relative magnitudes of **u** and **v**, can have a considerable effect on both the magnitude and direction of **p**. This can lead to difficulties when working with the cosine similarity measure, which is itself insensitive to the magnitudes of vectors. For example, if vector definitions are optimized by comparing the

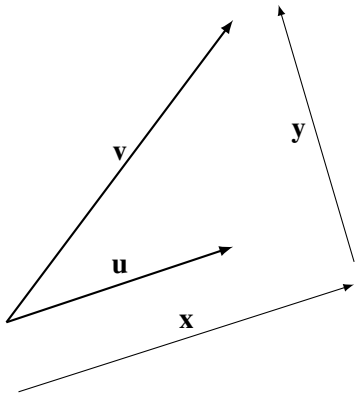


Figure 4. Vector **v** is decomposed into **x**, a component parallel to **u** and **y**, a component orthogonal to **u**.

predictions from the cosine similarity measure to some gold standard, then it is the directions of the vectors which are optimized, not their magnitudes. Utilizing vector addition as the composition function makes the product of the composition dependent on an aspect of the vectors which has not been optimized, namely their magnitude. Multiplicative combinations avoid this problem, because effects of the magnitudes of the constituents only show up in the magnitude of the product, which has no effect on the cosine similarity measure.

The multiplicative class of functions also allows us to think of one representation as modifying the other. This idea is fundamental in logic-based semantic frameworks (Montague, 1974) where different syntactic structures are given different function types. To see how the vector **u** can be thought of as something which modifies **v**, consider the partial product of **C** with **u**, producing a matrix which we shall call **U**.

$$\mathbf{p} = \mathbf{Cuv} = \mathbf{Uv} \tag{19}$$

Here, the composition function can be thought of as the action of a matrix, **U**, representing one constituent, on a vector, **v**, representing the other constituent. This is essentially Clark et al.'s (2008) approach to adjective-noun composition. In their scheme, nouns would be represented by vectors and adjectives by matrices which map the original noun representation to the modified representation. In our approach all syntactic types are simply represented by vectors; nevertheless, we can make use of their insight. Equation (19) demonstrates how a multiplicative composition tensor, C, allows us to map a constituent vector, **u**, onto a matrix, **U**, while representing all words with vectors.

Putting the simple multiplicative model (see equation (12)) into this form yields a matrix, **U**, whose off-diagonal elements are zero and whose diagonal elements are equal to the components of **u**.

$$U_{ij} = 0, U_{ii} = u_i \tag{20}$$

The action of this matrix on **v** is a type of dilation, in that it stretches and squeezes **v** in various directions. Specifically, **v** is scaled by a factor of u_i along the ith basis.

One drawback of this process is that its results are dependent on the basis used. Ideally, we would like to have a basis independent composition, i.e., one which is based solely on the geometry

of **u** and **v**.[7] One way to achieve basis independence is by dilating **v** along the direction of **u**, rather than along the basis directions. We thus decompose **v** into a component parallel to **u** and a component orthogonal to **u**, and then stretch the parallel component to modulate **v** to be more like **u**. Figure 4 illustrates this decomposition of **v** where **x** is the parallel component and **y** is the orthogonal component. These two vectors can be expressed in terms of **u** and **v** as follows:

$$\mathbf{x} = \frac{\mathbf{u}\cdot\mathbf{v}}{\mathbf{u}\cdot\mathbf{u}}\mathbf{u} \tag{21}$$

$$\mathbf{y} = \mathbf{v} - \mathbf{x} = \mathbf{v} - \frac{\mathbf{u}\cdot\mathbf{v}}{\mathbf{u}\cdot\mathbf{u}}\mathbf{u} \tag{22}$$

Thus, if we dilate **x** by a factor λ, while leaving **y** unchanged, we produce a modified vector, $\mathbf{v}'$, which has been stretched to emphasize the contribution of **u**:

$$\mathbf{v}' = \lambda\mathbf{x} + \mathbf{y} = \lambda\frac{\mathbf{u}\cdot\mathbf{v}}{\mathbf{u}\cdot\mathbf{u}}\mathbf{u} + \mathbf{v} - \frac{\mathbf{u}\cdot\mathbf{v}}{\mathbf{u}\cdot\mathbf{u}} = (\lambda-1)\frac{\mathbf{u}\cdot\mathbf{v}}{\mathbf{u}\cdot\mathbf{u}}\mathbf{u} + \mathbf{v} \tag{23}$$

However, since the cosine similarity function is insensitive to the magnitudes of vectors, we can multiply this vector by any factor we like without essentially changing the model. In particular, multiplying through by $\mathbf{u}\cdot\mathbf{u}$ makes this expression easier to work with:

$$\mathbf{p} = (\mathbf{u}\cdot\mathbf{u})\mathbf{v} + (\lambda-1)(\mathbf{u}\cdot\mathbf{v})\mathbf{u} \tag{24}$$

In order to apply this model to our example vectors, we must first calculate the dot products $\mathbf{practical}\cdot\mathbf{practical} = 156$ and $\mathbf{practical}\cdot\mathbf{difficulty} = 96$. Then, assuming λ is 2, the result of the composition is $96\ \mathbf{difficulty} + 156\ \mathbf{practical} = [96\ \ 1704\ \ 696\ \ 1944\ \ 624]$. This is now an asymmetric function of **u** and **v**, where **v** is stretched by a factor λ in the direction of **u**. However, it is also a more complex type of function, being quadratic in **u** (equation (7)).

Again, we can think of the composition of **u** with **v**, for this function (equation (24)), in terms of a matrix **U** which acts on **v**.

$$U_{i,j} = (\lambda-1)u_i u_j \tag{25}$$

$$U_{i,i} = (\sum_k u_k u_k) + (\lambda-1)u_i u_i \tag{26}$$

Where i, j and k range over the dimensions of the vector space.

The matrix **U** has one eigenvalue which is larger by a factor of λ than all the other eigenvalues, with the associated eigenvector being **u**. This corresponds to the fact that the action of this matrix on **v** is a dilation which stretches **v** differentially in the direction of **u**. Intuitively, this seems like an appropriate way to try to implement the idea that the action of combining two words can result in specific semantic aspects becoming more salient.

Collecting Similarity Ratings for Phrases

Vector-based models aimed at representing the meaning of individual words are commonly evaluated against human similarity judgments. For example, the benchmark dataset collected by Rubenstein and Goodenough (1965) consists of 65 noun-pairs ranging from highly synonymous (*gem-jewel*) to semantically unrelated (*noon-string*). For each pair, subjects gave similarity ratings

[7] This would allow, for example, the same composition function to be applied both to original vectors and to dimensionality reduced versions, without worrying about how to match the bases of these two spaces.

(on a scale of 0 to 4) whose average represents an estimate of the perceived similarity of the two words. Analogously, to evaluate the different composition models introduced above, we first had to elicit similarity judgments for phrases. Although such elicitation studies are less common in the literature, there is evidence that humans can reliably judge whether any two phrases are similar.

For example, Lapata and Lascarides present an experiment where participants rate whether adjective-noun combinations and their paraphrases have similar meanings, whereas other work (Lapata & Lascarides, 2003; Li, McLean, Bandar, O'Shea, & Crockett, 2006; Mitchell & Lapata, 2008) elicits similarity judgments for sentence pairs. In all cases, humans agree in their ratings, although the agreement tends to be lower compared to ratings assigned to isolated word pairs. Moreover, in computational linguistics, similarity judgments for phrases and sentences are routinely obtained as a means to evaluate the ability of an automatic system to generate paraphrases. Specifically, paraphrase pairs are presented to judges who are asked to decide whether they are semantically equivalent, i.e., whether they can be generally substituted for one another in the same context without great information loss (Barzilay & Lee, 2003; Bannard & Callison-Burch, 2005). Participants are usually asked to rate the paraphrase pairs using a nominal scale (e.g., definitely similar, sometimes similar, never similar).

In our experiments, we collected similarity judgments for adjective-noun, noun-noun, and verb-object combinations using a rating scale. Following previous work (Bullinaria & Levy, 2007; Padó & Lapata, 2007; McDonald, 2000), we then used correlation analysis to examine the relationship between the human ratings and their corresponding vector-based similarity values. In this section we describe our method for assembling the set of experimental materials and eliciting similarity ratings for these stimuli.

Materials and Design

We evaluated the predictions of our composition models against similarity ratings which we obtained for three types of phrases, adjective-nouns, noun-nouns, and verb-objects. These phrases were selected from the British National Corpus (BNC), as we wanted to ensure they represented genuine usage. We also chose relatively frequent phrases to avoid confounding effects from infrequent or unfamiliar constructions. Ideally, we would also like our phrase pairs to be representative of the full variation in semantic similarity. A potential caveat here concerns the difficulty of the task. For example, requiring fine grained similarity judgments would fail to yield ratings with robust inter-subject agreement, a prerequisite for reliably evaluating the models. On the other hand, the relatively easy task of simply discriminating high similarity from low similarity items, will result in all models achieving high scores and thus failing to clearly distinguish their performance.

Our approach was to collect pairs representative of three similarity "bands" (High, Medium, and Low) — applying a simple method that randomly samples and pairs phrases from a corpus yields items that are mostly semantically unrelated, with only a few showing weak similarity. Initially, we extracted all adjective-noun, noun-noun, and verb-object combinations attested in the corpus. The latter was parsed with RASP (Briscoe & Carroll, 2002), a broad coverage syntactic analyzer. Our high similarity items were compiled from phrases occurring at least 100 times in the BNC. For each grammatical construction, any two phrases were considered highly similar if swapping their heads resulted into two new phrases which were also attested in the BNC at least 100 times. For example, *practical difficulty–economic problem* is a candidate high similarity item, because *practical problem* and *economic difficulty* are also high frequency phrases. Our hypothesis was that the phrases resulting from this recombination process must exhibit some semantic

overlap, especially if they appear often in the BNC. This procedure resulted in 11,476 candidate adjective-noun, 366 noun-noun, and 1,004 verb-object pairs.

In order to reduce the set of items to a more manageable size and more importantly to guarantee that the phrases were indeed semantically similar, we resorted to WordNet (Fellbaum, 1998). We used a well-known dictionary-based similarity measure, originally proposed by Lesk (1986), to rank the candidate phrase pairs. According to this measure, the semantic relatedness of two words is proportional to the extent of overlap of their dictionary definitions[8] (*glosses* in WordNet). We computed the similarity of two phrases, as the sum of the similarities of their constituents. The 36 highest ranking phrase pairs (for each grammatical structure) on this measure formed our high-similarity items (e.g., *vast amount–large quantity*, *telephone number–phone call*, *start work–begin career*). These 36 phrase pairs (72 phrases in total) were subsequently recombined to produce the items in the medium and low similarity bands. This was done in order to eliminate any confounding effects relating to the vocabulary of the individual phrases. By choosing the same same set of phrases to construct all three bands, differences between bands cannot be attributed to lexical choice but instead to their actual similarity relations.

Specifically, the high similarity phrases were first randomly split into the three groups, and then candidate items for the remaining bands were constructed by pairing phrases from each of these groups. So, each phrase was used three times in our materials: once in a high similarity pair, once in a medium pair and once in a low pair. For example, *practical difficulty* from the first group was paired with *effective way* from the third group to produce the item *practical difficulty–effective way*. The Lesk similarity for each of these pairs was calculated as above and the 36 highest ranking items on this measure were selected, subject to the constraint that each phrase was only used once in each group. This produced a set of Medium similarity items, which, while they scored reasonably highly on the WordNet-based measure, did not have the recombination property described above (e.g., *social activity–economic condition*, *market leader–board member*, *discuss issue–present problem*). A further 36 items were selected from the same set of candidate items, though in this case by choosing the lowest ranking items. This produced a set of Low similarity items (e.g., *practical difficulty–cold air*, *phone call–state benefit*, *drink water–use test*). The entire list of experimental stimuli is given in Appendix A.

Thus, in our experimental design, the subject ratings and model predictions were the dependent variables, and the bands and groups acted as blocking factors with a 3×3 structure. For each phrase type (i.e., adjective-noun, noun-noun, and verb-object) we collected 108 items, 12 for each band by group cell. The selected verb-object pairs were converted into a simple sentence by adding a subject and articles or pronouns where appropriate. All verbs were in the past tense. The sentential subjects were familiar proper names (BNC corpus frequency > 30 per million) balanced for gender.

Procedure

The elicitation studies were conducted online using Webexp (Keller, Gunasekharan, Mayo, & Corley, 2009), an interactive software package for administering web-based psychological experiments. Subjects took part in an experimental session that lasted approximately 20 minutes. The experiment was self-paced, and response times were recorded to allow the data to be screened for anomalies. Subjects accessed the experiment using their web browser, which established an Internet connection to the experimental server running WebExp 2.1

[8] We used the implementation provided in the WordNet Similarity package (Pedersen, Patwardhan, & Michelizzi, 2004).

Table 1: Descriptive statistics for similarity experiments (adjective-noun, noun-noun, and verb-object), by subjects.

	Adjective-Noun			Noun-Noun			Verb-Object		
	Mean	SD	SE	Mean	SD	SE	Mean	SD	SE
High	3.76	1.926	0.093	4.13	1.761	0.085	3.91	2.031	0.098
Medium	2.50	1.814	0.087	3.04	1.732	0.083	2.85	1.775	0.085
Low	1.99	1.353	0.065	2.80	1.529	0.074	2.38	1.525	0.073

Subjects were given instructions that explained the task and provided examples (our instructions for the adjective-noun similarity experiment are reproduced in Appendix B). They were asked to judge the similarity of phrases using a seven point rating scale where a high number indicates higher similarity. To familiarize subjects with the similarity rating task, the experiment consisted of a practice phase (of five items), followed by the experimental phase. In both phases, the participants saw one phrase pair at a time and rated its similarity by clicking on one of seven buttons displaying the numbers 1 to 7. The set of practice and experimental items was presented in random order.

Subjects

The experiment was completed by unpaid volunteers, all self-reported native speakers of English. Subjects were recruited by postings to local email lists; they had to be linguistically naive, neither linguists nor students of linguistics were allowed to participate. The adjective-noun experiment was completed by 88 participants; 69 subjects took part in the noun-noun experiment and 91 in the verb-object experiment. 14 participants were eliminated because they were non-native English speakers. The data of 30 subjects was excluded after inspection of their responses revealed anomalies in their ratings. For example, they were pressing buttons randomly, alternately, or rated all phrase pairs uniformly. This left 204 subjects for analysis, 72 for the adjective-noun, 56 for the noun-noun, and 76 for the verb-object experiment. 35 participants were male and 73 female, 94 were right-handed, and 14 left-handed. The subject ages ranged from 17 to 66, the mean was 31. Participants were randomly allocated to a development set, used for optimizing model parameters, and a test set on which the final evaluation of all models was carried out. For each experiment the test set contained 36 participants, and the development set contained 18.

Results

We first performed a series of Kruskal-Wallis rank sum tests to examine the relationship between our similarity bands and the elicited similarity ratings. Within each experiment, the subject ratings were significantly different ($p < 0.01$) across all bands, and also between each pair of bands. Furthermore, the statistics in Table 1 demonstrate that the mean ratings show the correct ordering ($High > Medium > Low$) and that there is substantial overlap between each band. These results confirm that our procedure for generating the materials produced items with a wide range of similarities.

We further examined how well the participants agreed in their similarity judgments for adjective-noun, noun-noun, and verb-object combinations. Inter-subject agreement gives an upper bound for the task and allows us to interpret how well our models are doing in relation to humans.

To calculate inter-subject agreement we used leave one-out resampling. The technique is a special case of n-fold cross-validation (Weiss & Kulikowski, 1991) and has been previously used for measuring how well humans agree on judging semantic similarity (Resnik & Diab, 2000; Resnik, 1999). For each subject group we divided the set of the subjects' responses with size m into a set of size $m-1$ (i.e., the response data of all but one subject) and a set of size one (i.e., the response data of a single subject). We then correlated the ratings of the former set with the ratings of the latter using Spearman's ρ correlation coefficient. This was repeated m times. For the adjective-noun experiment, the average inter-subject agreement was .52 (Min $= 0.35$, Max $= 0.73$, SD $= 0.12$), for the noun-noun experiment .51 (Min $= 0.36$, Max $= 0.58$, SD $= 0.06$), and for the verb-object experiment 0.55 (Min $= 0.45$, Max $= 0.65$, SD $= 0.06$). These results indicate that the participants found the similarity rating task relatively difficult, though still produced ratings with a reasonable level of consistency.

Modeling Experiments

Semantic Representation

Irrespectively of their form, all composition models discussed here rely on vector-based representations for individual words. Our experiments examined two such representations and their impact on composition. Our first model is a simple and popular (McDonald, 2000; Bullinaria & Levy, 2007; Lowe, 2000) semantic space that associates each vector component with a particular context word, and assigns it a value based on the strength of its co-occurrence with the target (i.e., the word for which a semantic representation is being constructed). This model has the benefits of simplicity and also of being largely free of any additional theoretical assumptions over and above the distributional approach to semantics.

For our experiments, we built the semantic space on a lemmatized version of the BNC. Following previous work (Bullinaria & Levy, 2007), we optimized its parameters on a word-based semantic similarity task. The task involves examining the degree of correlation between the human judgments for two individual words and vector-based similarity values. We experimented with a variety of dimensions (ranging from 50 to 500,000), vector component definitions (e.g., pointwise mutual information or log likelihood ratio) and similarity measures (e.g., cosine or confusion probability). We used WordSim353, a benchmark dataset (Finkelstein et al., 2002), consisting of relatedness judgments (on a scale of 0 to 10) for 353 word pairs.

We obtained best results with a model using a context window of five words on either side of the target word and 2,000 vector components. The latter were the most common context words (excluding a list of stop words). These components were set to the ratio of the probability of the context word given the target word to the probability of the context word overall:

$$v_i(t) = \frac{p(c_i|t)}{p(c_i)} = \frac{freq_{c_i,t} \cdot freq_{total}}{freq_t \cdot freq_{c_i}} \tag{27}$$

Where $freq_{c_i,t}$, $freq_{total}$, $freq_t$ and $freq_{c_i}$ are the frequencies of the context word c_i with the target word t, the total count of all word tokens, the frequency of the context word c_i and the frequency of the target word t, respectively.

This configuration gave high correlations with the WordSim353 similarity judgments using the cosine measure ($\rho = 0.42$). In addition, Bullinaria and Levy (2007) found that these parameters perform well on a number of other tasks such as the synonymy task from the *Test of English as*

a Foreign Language (TOEFL). We compute the similarity between two vectors $\mathbf{v}(t_1)$ and $\mathbf{v}(t_2)$ representing target words t_1 and t_2, respectively as:

$$sim(t_1,t_2) = cos(\mathbf{v}(t_1),\mathbf{v}(t_2)) = \frac{\mathbf{v}(t_1)\cdot\mathbf{v}(t_2)}{|\mathbf{v}(t_1)||\mathbf{v}(t_2)|} \tag{28}$$

Probabilistic topic models offer an alternative to semantic spaces. Although several variants have been proposed in the literature (e.g., Griffiths et al., 2007; Blei et al., 2003) they are all based on the same fundamental idea: documents are mixtures of topics where a topic is a probability distribution over words. And the content of a topic is expressed by the probabilities of the words within that topic. A topic model is a *generative* model specifying a specific process of how to generate a document. Our experiments are based on the Latent Dirichlet Allocation (LDA, Blei et al., 2003) topic model where the generative process for a document d is as follows. We first draw the mixing proportion over topics θ_d from a Dirichlet prior[9] with parameters α. Next, for each of the N_d words w_{dn} in document d, a topic z_{dn} is first drawn from a multinomial distribution with parameters θ_{dn}. The probability of a word token w taking on value i given that topic $z = j$ is parametrized using a matrix β with $b_{ij} = p(w = i|z = j)$. Integrating out θ_d's and z_{dn}'s, gives $P(D|\alpha,\beta)$, the probability of a corpus (or document collection):

$$\prod_{d=1}^{M}\int P(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}}P(z_{dn}|\theta_d)P(w_{dn}|z_{dn},\beta)\right)d\theta_d \tag{29}$$

The central computational problem in topic modeling is to obtain the posterior distribution $P(\theta,\mathbf{z}|\mathbf{w},\alpha,\beta)$ of the hidden variables given a document $\mathbf{w} = (w_1,w_2,\ldots,w_N)$. Although this distribution is intractable in general, a variety of approximate inference algorithms have been proposed in the literature including expectation maximization, variational expectation maximization, expectation propagation, several forms of Markov chain Monte Carlo (MCMC), and variational inference. Our model adopts the Gibbs sampling procedure discussed in Griffiths et al. (2007).

Under this model, constructing a semantic representation for a target word amounts to estimating the topic proportions for that word. We therefore select the number of topics, K, and train the LDA algorithm on a document collection to obtain the β parameters, where β represents the probability of a word w_i given a topic z_j, $p(w_i|z_j) = \beta_{ij}$. The meaning of w_i is thus extracted from β and is a K-element vector, whose components correspond to the probability of w_i given each topic assumed to have generated the document collection. Figure 5 gives an example of the semantic representations extracted by the LDA model. Similarity in this model can be also measured as the cosine of the angle between the topic vectors representing any two words.

For our experiments, we trained an LDA model on the BNC corpus.[10] We optimized the model's parameters in terms of correlation on the same WordSim353 dataset used for the simpler semantic space model. We varied the number of topics from 10 to 1000 and obtained best results

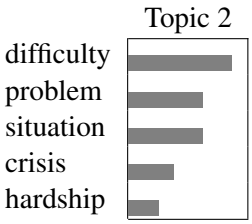
[9] The Dirichlet distribution is a commonly used prior for multinomials $P(\theta) = \frac{1}{B(a_1\ldots,a_n)}\prod_{i=1}^{n}\theta_i^{a_i-1}$ where $a_1\ldots,a_n$ are the parameters of the prior and the normalizing constant $B(a_1\ldots,a_n)$ is the n-dimensional Beta function. One important reason for the use of the Dirichlet prior in the case of multinomial parameters is its mathematical expedience. It is a *conjugate prior*, i.e., of the same functional from as the likelihood function for the multinomial. This means that the prior and the likelihood can easily combine according to Bayes' law to specify the posterior distribution $P(\theta|c_1\ldots,c_k)$ where $c_1\ldots,c_k$ are the counts for each outcome.

[10] The implementation we used is available at `http://gibbslda.sourceforge.net/`.

Figure 5. Semantic representations obtained from the LDA topic model

	Topic 1	Topic 2	Topic 3	Topic *n*
practical	0.39	0.02	0.32	...
difficulty	0.03	0.44	0.11	...
produce	0.06	0.17	0.09	...

(a) Meaning representations obtained from the LDA model; rows are target words, columns correspond to $P(w|z)$, the probability of a target word w given a topic z.



(b) Topic 2 represented as a probability distribution over words (only the five words that have the highest probability under this topic are shown.

Table 2: Parameters for Kintsch's composition model.

	Semantic Space		LDA	
	m	*k*	*m*	*k*
Adjective-Noun	10	10	50	10
Noun-Noun	100	1	50	1
Verb-Object	50	10	100	1

with 100 topics ($\rho = 0.48$). The hyperparameters α and β were initialized to 0.1, and 0.01 respectively.

Parameters for Composition Models

Using the semantic representations described above, our experiments assessed the performance of the simple additive and multiplicative models (see equations (8) and (12), respectively), Kintsch's (2001) model, the tensor product (equation (14)) and circular convolution (equation (17)). Note that Kintsch's model has two free parameters, the m neighbors most similar to the head, and the k of m neighbors closest to its dependent, which we optimized on the development set. Table 2 shows the best parameters for the semantic space and LDA topic models, respectively. In addition to these models, we also considered two models based on the weighted sum of two values. These were the weighted addition model (equation (10)) and the dilation model (equation (24)). We tuned the parameters for these models on the development set. These include the weights α and β for the additive models, the dilation factor λ for the dilation models and their direction[11]. The optimal parameters for these models are shown in Tables 3 and 4.

Finally, as a baseline, we considered two non-compositional models. The first simply uses the vector representing the head of the phrase as the representation of the whole phrase (equation (11)), and the second treats the phrase as a single target unit and thus extracts a vector representation for the whole phrase. The latter baseline is only applicable to the standard semantic space. LDA derives semantic representations for individual words rather than word combinations (although extensions

[11]The dilation model can be applied in two ways, since the function is asymmetric in $\mathbf{u}$ and $\mathbf{v}$. Either $\mathbf{u}$ can be used to dilate $\mathbf{v}$, or $\mathbf{v}$ can be used to dilate $\mathbf{u}$.

Table 3: Parameters for weighted addition composition models.

	Semantic Space		LDA	
	α	β	α	β
Adjective-Noun	0.88	0.12	0.65	0.35
Noun-Noun	0.32	0.68	0.34	0.66
Verb-Object	0.31	0.69	0.50	0.50

Table 4: Parameters for dilation models.

	Semantic Space		LDA	
	λ	Direction	λ	Direction
Adjective-Noun	16.7	Adjective	2.2	Noun
Noun-Noun	8.3	Head Noun	7.1	Head Noun
Verb-Object	7.7	Verb	6.3	Verb

of the basic model have been proposed that take word order into account; see Wallach (2002) for an example). Table 5 gives the details of the composition functions we evaluated, expressed in terms of the vector components for each model.

Evaluation

We evaluated the proposed composition models via correlation analysis. Specifically, the elicited similarity ratings were correlated with our models' predictions using Spearman's ρ correlation coefficient.[12] Given some composition function, $f(\cdot,\cdot)$, and two phrases a_1b_1 and a_2b_2, we applied f to the vectors $\mathbf{u_1}$ and $\mathbf{v_1}$ representing a_1 and b_1, respectively, to produce a composite representation, $\mathbf{p_1}$. Analogously, vectors $\mathbf{u_2}$ and $\mathbf{v_2}$ yield $\mathbf{p_2}$ as a representation for a_2b_2. Under this setup, we can calculate the similarity of two phrases by measuring their distance in semantic space. A large number of such measures have been proposed in the literature (see Bullinaria and Levy (2007) and Weeds (2003) for an overview). We opted for the widely used cosine measure (see equation (28)) due to its simplicity and good performance in simulating word similarity ratings (Bullinaria & Levy, 2007; McDonald, 2000; Griffiths et al., 2007).

Results

Table 6 shows the correlation of the subjects' similarity ratings with the models' predictions when using a simple co-occurrence-based semantic space. All models are significantly correlated with the human judgments ($p < 0.01$). The only exception is circular convolution when applied to noun-noun combinations. Let us first consider, the simpler composition models based on vector addition (see Additive and Kintsch in the table). Within this class of models we observe that Kintsch's model fails to improve on the simple additive model and is significantly[13] worse ($p < 0.01$) than the

[12] We avoided correlating the model predictions with averaged participant judgments as this is inappropriate given the ordinal nature of the scale of these judgments and also leads to a dependence between the number of participants and the magnitude of the correlation coefficient.

[13] We examined whether the correlations achieved differ significantly using a t-test (Cohen & Cohen, 1983).

Table 5: Composition functions considered in our experiments.

Model	Function
Additive	$p_i = u_i + v_i$
Kintsch	$p_i = u_i + v_i + n_i$
Multiplicative	$p_i = u_i \cdot v_i$
Tensor Product	$p_{i,j} = u_i \cdot v_j$
Circular Convolution	$p_i = \sum_j u_j \cdot v_{i-j}$
Weighted Additive	$p_i = \alpha v_i + \beta u_i$
Dilation	$p_i = v_i \sum_j u_j u_j + (\lambda - 1) u_i \sum_j u_j v_j$
Head Only	$p_i = v_i$
Target Unit	$p_i = v_i (t_1 t_2)$

Table 6: Correlation coefficients of model predictions with subject similarity ratings (Spearman's ρ) using a simple semantic space.

Model	Adjective-Noun	Noun-Noun	Verb-Object
Additive	0.36	0.39	0.30
Kintsch	0.32	0.22	0.29
Mutliplicative	0.46	0.49	0.37
Tensor Product	0.41	0.36	0.33
Convolution	0.09	0.05	0.10
Weighted Additive	0.44	0.41	0.34
Dilation	0.44	0.41	0.38
Target Unit	0.43	0.34	0.29
Head Only	0.43	0.17	0.24
Humans	0.52	0.49	0.55

standard additive model for the noun compounds.

Within the class of multiplicative models (see Multiplicative, Tensor Product, and Circular Convolution in Table 6), the simple multiplicative model significantly ($p < 0.01$) outperforms all other models. Specifically, both tensor products and circular convolution are significantly worse ($p < 0.01$). The multiplicative model is also significantly better than the Additive one ($p < 0.01$). These results are observed across the board, with adjective-noun, noun-noun, and verb-object combinations. It is worth noting that circular convolution is the worst performing model. The tensor product itself, from which circular convolution is derived, is significantly better ($p < 0.01$) in all experiments. This indicates that the manner in which circular convolution projects the tensor product down onto a lower dimensional space does not preserve any useful information the product may have contained. In addition, the fact that the tensor product is significantly worse than the simple multiplicative model indicates that the off diagonal elements of the product, which are discarded in the simple multiplicative model, are probably not contributing much to the composition.

We next consider the Weighted Additive and Dilation models. Recall that these models are parametrized; in dilation models the modifier dilates the head by a factor λ whereas the weighted

Table 7: Correlation coefficients of model predictions with subject similarity ratings (Spearman's ρ) using the LDA topic model.

Model	Adjective-Noun	Noun-Noun	Verb-Object
Additive	0.37	0.45	0.40
Kintsch	0.30	0.28	0.33
Multiplicative	0.25	0.45	0.34
Tensor Product	0.39	0.43	0.33
Convolution	0.15	0.17	0.12
Weighted Additive	0.38	0.46	0.40
Dilation	0.38	0.45	0.41
Head Only	0.35	0.27	0.17
Humans	0.52	0.49	0.55

additive model weights the constituents in the summation differentially. As shown in Table 6 the two models perform similarly. This is not entirely surprising, as both consist of a sum of the constituents multiplied by scalar factors (see equations (10) and (24)). The performance of these models does not differ significantly, except in the case of verb-object combinations where the dilation model performs significantly better ($p < 0.01$). We conjecture that the dilation model is more accurate at capturing selectional restrictions. This model also fares similarly to the multiplicative model. The two models yield correlations that are not significantly different, except in the case of noun-noun combinations, where the multiplicative model is better ($p < 0.01$).

The two non-compositional models, Target Unit and Head Only, perform worse than multiplicative composition, with this difference reaching significance ($p < 0.01$) for noun-noun and verb-object combinations. In general, the target unit model performs better than the head only model (it obtains significantly ($p < 0.01$) better correlations for noun-noun combinations). This is not surprising, the target unit model may be non-compositional, but nevertheless represents the semantics of the two words participating in the composition more faithfully, whereas the head only model offers a more impoverished representation as it is based solely on the meaning of the head.

In sum, we find that the multiplicative, weighted additive and dilation models perform overall best. The multiplicative model has a slight advantage as it has no parameters (other than the semantic space representing the individual words), and is conceptually simpler than the other two models. On the down side, it is does not take syntactic information into account, whereas the other two can modulate the role of syntactic structure by tuning the appropriate weights. We should also note that in all cases our compositional models fall behind the human upper bound (see the last row in Table 6). The multiplicative model comes close when applied to noun-noun combinations.

We now turn our attention to the compositional models which employ the LDA topic model. As can be seen in Table 7, Kintsch's model remains worse than the simple additive model for all constructions considered here (and the differences are statistically significant ($p < 0.01$)). Regarding compositional models based on multiplication, we observe that tensor products and the simple multiplicative model yield comparable performances for noun-noun and verb-object combinations. They differ for adjective-nouns with the tensor product being significantly better ($p < 0.01$). Circular convolution remains the worst performing model. Not surprisingly, Weighted Additive and Dilation models obtain almost identical performances. And they are not significantly different from

the simple Additive model. The non-compositional model (Head Only) is significantly worse than these models. Comparing the spatial and topic-based representations reveals that the multiplicative composition model on the simple semantic space is significantly ($p < 0.01$) better than the dilation model with LDA, except in the verb-object experiment, where there is no significant difference between them.

In conclusion, we observe that dilation models perform consistently well across representations. This is not entirely unexpected as they are more flexible than other compositional models due to their parametric nature. They can be tuned to model more faithfully specific syntactic constructions while being sensitive to the underlying semantic representation. Our results also indicate that additive composition functions work best with the LDA topic model, whereas a multiplicative composition function produced the most predictive similarity values with a simple semantic space. We attribute the disparity in performance to the sparsity of the LDA representations. The simple semantic space contains highly distributed representations, with the semantic content spread across the great variety of contexts a target word occurs in. In contrast, topic models tend to produce representations in which the vast majority of topics are inactive (i.e., zero) and when these topics are multiplied by other topics, the result is zero. Thus, multiplicative combinations of sparse representations tend to result in a loss of useful information.

Discussion

In this paper we presented a framework for vector-based semantic composition. We formulated composition as a function of two vectors and introduced several models based on addition and multiplication. These models were applied to vectors corresponding to distinct meaning representations: a simple semantic space based on word co-occurrences and a topic-based model built using LDA. We compared the model predictions empirically on a phrase similarity task, using ratings elicited from native speakers. Overall, we observe that dilation models perform consistently well across semantic representations. A compositional model based on component-wise multiplication performs best on the simple semantic space, whereas additive models are preferable with LDA. Interestingly, we also find that the compositional approach to constructing representations outperforms a more direct non-compositional approach based on treating the phrases essentially as single lexical units. This is not entirely surprising as our materials which were compiled so as to avoid a high degree of lexicalization. Such an approach may be better suited to modeling non-compositional structures that are lexicalized and frequently occurring (Baldwin et al., 2003; Bannard et al., 2003).

Despite this success, a significant weakness of many of the models considered here is their insensitivity to syntax. The multiplicative model, in particular, is symmetric, and thus makes no distinction between the constituents it combines. Yet, in spite of this, it is the strongest model for the simple semantic space. And although the weighted addition and dilation models differentiate between constituents, their dependence on syntax is rather limited, involving only a differential weighting of the contribution of each constituent. Perhaps more importantly, none of the representations could be said to have any internal structure. Thus, they cannot be broken down into parts which can be independently interpreted or operated upon. Symbolic representations, in contrast, build complex structures by, for example, binding predicates to arguments. In fact, it is often argued that however composition is implemented it must exhibit certain features characteristic of this symbolic binding (Fodor & Pylyshyn, 1988; Holyoak & Hummel, 2000).

Our results do not indicate that models which mimic symbolic binding (i.e., tensor products and circular convolution) are better than those that don't (at least for the phrase similarity task and

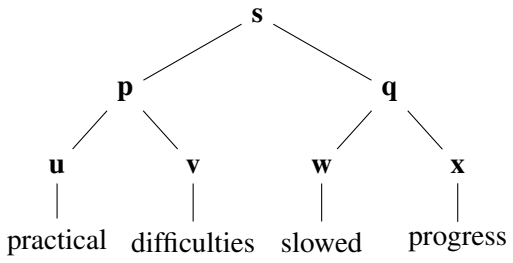


Figure 6. Example of composition operating over parse trees.

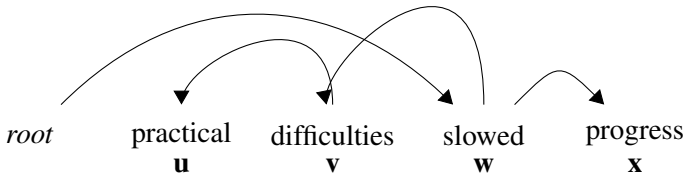


Figure 7. Example of composition operating over dependency structures.

the syntactic structures we examined). In particular, circular convolution is, across the board, the worst performing model. One issue in the application of circular convolution is that it is designed for use with random vectors, as opposed to the structured semantic vectors we assume here. A more significant issue, however, concerns symbol binding in general which is somewhat distinct from semantic composition. In modeling the composition of an adjective with a noun, it is not enough to simply bind the representation of one to the representation of the other, we must instead model the interaction between their meanings and their integration to form a whole. Circular convolution is simply designed to allow a pair of vectors to be bound in a manner that allows the result to be decomposed into its original constituents at a later time. This may well be adequate as a model for syntactic operations on symbols, but, as our results show, it is not, by itself, enough to model the process of semantic composition. Nevertheless, we anticipate further improvements to our vector-based composition models will involve taking a more sophisticated approach to the structure of representations, in particular with regard to predicate-argument structures. Our results also suggest that assuming a single semantic representation may not be sufficient for all tasks. For instance, it is not guaranteed that the same highly structured representations appropriate for deductive inference will also provide a good model for semantic similarity. Semantics, covering such a wide range of cognitive phenomena, might well be expected to involve multiple systems and processes, which make use of quite distinct representations.

In this article, we have been concerned with modeling the similarity between simple phrases, consisting of heads and their dependents. We have thus avoided the important question of how vectors compose to create representations for larger phrases and sentences. It seems reasonable to

assume that the composition process operates over syntactic representations such as binary parse trees. A sentence will typically consist of several composition operations, each applied to a pair of constituents **u** and **v**. Figure 5 depicts this composition process for the sentence *practical difficulties slowed progress*. Initially, *practical* and *difficulties* are composed into **p**, and *slowed* and *progress* into **q**. The final sentence representation, **s**, is the composition of the pair of phrase representations **p** and **q**. Alternatively, composition may operate over dependency graphs representing words and their relationship to syntactic modifiers using directed edges (see the example in Figure 6).

It is interesting then to consider which composition function would be best suited for representing sentences. For example, we could adopt different functions for different constructions. Our experiments show that the simple multiplicative model performs best at modeling adjective-noun and noun-noun combinations, whereas the dilation model is better for verb-object constructions. Alternatively, we could adopt a single composition function that applies uniformly across all syntactic relations. As discussed earlier, the simple multiplicative function is insensitive to syntax and word order. The dilation model, however, remedies this. It is also based on a multiplicative composition function, but can take syntax into account by stretching one vector along the direction of another one (see equation (24)).

Overall, we anticipate that more substantial correlations with human similarity judgments can be achieved by implementing more sophisticated models from within the framework outlined here. In particular, the general class of multiplicative models (see equation (6)) appears to be a fruitful area to explore. Future directions include constraining the number of free parameters in linguistically plausible ways and scaling to larger datasets. The applications of the framework discussed here are many and varied. We intend to assess the potential of our composition models on context sensitive semantic priming (Till, Mross, & Kintsch, 1988) and inductive inference (Heit & Rubinstein, 1994). Another interesting application concerns sentence processing and the extent to which the compositional models discussed here can explain reading times in eye-tracking corpora (Pynte, New, & Kennedy, 2008; Demberg & Keller, 2008).

References

Aerts, D., & Czachor, M. (2004). Quantum aspects of semantic analysis and symbolic artificial intelligence. *Journal of Physics A-Mathematical and General*, *37*, L123–L32.

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review*(3), 463–498.

Baldwin, T., Bannard, C., Tanaka, T., & Widdows, D. (2003). An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions* (pp. 89–96). Morristown, NJ.

Bannard, C., Baldwin, T., & Lascarides, A. (2003). A statistical approach to the semantics of verb-particles. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions* (pp. 65–72). Morristown, NJ.

Bannard, C., & Callison-Burch, C. (2005). Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd annual meeting of the association for computational linguistics* (pp. 597–604). Ann Arbor.

Barzilay, R., & Lee, L. (2003). Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of the human language technology conference and annual meeting of the north american chapter of the association for computational linguistics* (pp. 16–23).

Blackburn, P., & Bos, J. (2005). *Representation and inference for natural language: A first course in computational semantics.* Seattle, WA: Stanford: CSLI Press.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Briscoe, E., & Carroll, J. (2002). Robust accurate statistical annotation of general text. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation* (pp. 1499–1504). Las Palmas, Canary Islands.

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

Clark, S., Coecke, B., & Sadrzadeh, M. (2008). A compositional distributional model of meaning. In *Proceedings of the 2nd Symposium on Quantum Interaction* (pp. 133–140). Oxford, UK: College Publications.

Clark, S., & Pulman, S. (2007). Combining symbolic and distributional models of meaning. In *Proceedings of the AAAI Spring Symposium on Quantum Interaction, Stanford, CA, 2007* (pp. 52–55).

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences*. NJ: Erlbaum: Hillsdale.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*(8), 240-248.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, *41*(6), 391-407.

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

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *101*(2), 193–210.

Denhire, G., & Lemaire, B. (2004). A computational model of children's semantic memory. In *Proceedings of the 26th Annual Meeting of the Cognitive Science Society* (pp. 297–302). Mahwah, NJ: Lawrence Erlbaum Associates.

Dennis, S. (2007). Introducing word order in an lsa framework. In P. Press (Ed.), *Latent semantic analysis: A road to meaning* (pp. 449–464). Thomas K. Landauer and Danielle S. McNamara and Simon Dennis and Walter Kintsch.

Doumas, L. A. A., & Hummel, J. E. (2005). Modeling human mental representations: What works and what doesn't and why. In K. J. Holyoak & R. G. Morrison (Eds.), *The cambridge handbook of thinking and reasoning* (pp. 73–91). Cambridge, UK: Cambridge University Press.

Doumas, L. A. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, *115*, 1–43.

Duffy, S. A., Henderson, J. M., & Morris, R. K. (1989). Semantic facilitation of lexical access during sentence processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 791–801.

Eliasmith, C., & Thagard, P. (2001). Integrating structure and meaning: A distributed model of analogical mapping. *Cognitive Science*, *25*(1), –.

Elman, J., Bates, E., Johnson, M., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective*. Cambridge, MA: MIT Press/Bradford Books.

Estes, W. K. (1994). *Classification and cognition*. New York: Oxford University Press.

Fellbaum, C. (Ed.). (1998). *Wordnet: An electronic lexical database (language, speech, and communication)*. The MIT Press. Hardcover.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., et al. (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, *20*(1), 116–131.

Fodor, J., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*, 3-71.

Foltz, P. W., Kintsch, W., & Landauer, T. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Process*, *15*, 285–307.

Foss, D. J. (1982). A discourse on semantic priming. *Cognitive Psychology*, *14*, 590–607.

Frank, S., Koppen, M., Noordman, L., & Vonk, W. (2007). World knowledge in computational models of discourse comprehension. *Discourse Processes*. (In press)

Frege, G. (1884). *Die Grundlagen der Arithmetik*. Breslau: W. Koebner.

Gentner, D. (1989). The mechanisms of analogical learning. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 199–241). Cambridge: Cambridge University Press.

Grefenstette, G. (1994). *Explorations in automatic thesaurus discovery*. Norwell, MA, USA: Kluwer Academic Publishers.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211–244.

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

Heit, E., & Rubinstein, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 411–422.

Hinton, J., & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review*, *98*, 74–95.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, *42*(1-2), 177–196.

Holyoak, K. J., & Hummel, J. E. (2000). The proper treatment of symbols in a connectionist architecture. In E. Dietrich & A. Markman (Eds.), *Cognitive dynamics: Conceptual change in humans and machines.* (pp. 229–264). Cambridge, MA: MIT Press.

Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory and Cognition*, *15*, 332–340.

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

Kako, E. (1999). Elements of syntax in the systems of three language-trained animals. *Animal Learning and Behavior*, *27*, 1–14.

Kanerva, P. (1988). *Sparse distributed memory*. Cambridge, MA: MIT Press.

Kanerva, P. (2009). Hyperdimensional computing: An introduction to computing in distributed representation with high dimensional random vectors. *Cognitive Computation*, *1*, 139–159.

Kanerva, P., Kristoferson, J., & Holst, A. (2000). Random indexing of text samples for latent semantic analysis. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of proceedings of the 22nd annual conference of the cognitive science society* (p. 1036). Mahwah, New Jersey: Erlbaum.

Keller, F., Gunasekharan, S., Mayo, N., & Corley, M. (2009). Timing accuracy of web experiments: A case study using the WebExp software package. *Behavior Research Methods*, *41*(1), 1–12.

Kintsch, W. (1988, April). The role of knowledge in discourse comprehension: a construction-integration model. *Psychological Review*, *95*(2), 163–182.

Kintsch, W. (2001). Predication. *Cognitive Science*, *25*(2), 173–202.

Laham, D. R. (2000). *Automated content assessment of text using latent semantic analysis to simulate human cognition*. Unpublished doctoral dissertation, University of Colorado at Boulder.

Lakoff, G. (1977). Linguistic gestalts. In W. Beach, S. Fox, & S. Philosoph (Eds.), *Papers from the 13th Regional Meeting, Chicago Linguistic Society* (pp. 236–287). Chicago, Illinois: Chicago Linguistic Society.

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.

Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. E. (1997a). How well can passage meaning be derived without using word order? a comparison of latent semantic analysis and humans. In Erlbaum (Ed.), *Proceedings of the nineteenth annual conference of the cognitive science society* (pp. 412–417). Mahwah, NJ.

Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. E. (1997b). How well can passage meaning be derived without using word order: A comparison of latent semantic analysis and humans. In *Nineteenth annual conference of the cognitive science society* (pp. 412–417). Stanford, CA: Lawrence Erlbaum.

Lapata, M., & Lascarides, A. (2003). A probabilistic account of logical metonymy. *Computational Linguistics*, *29*(2), 263–317.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proc. of the 5th sigdoc* (pp. 24–26). New York, NY.

Li, Y., McLean, D., Bandar, Z., O'Shea, J., & Crockett, K. (2006). Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge and Data Engineering*, *18*(8), 1138–1149.

Lin, D. (1998). Automatic Retrieval and Clustering of Similar Words. In *Proceedings of the 17th International Conference on Computational Linguistics* (pp. 768–774).

Lowe, W. (2000). What is the dimensionality of human semantic space? In R. M. French & J. P. Sougné (Eds.), *Proceedings of the 6th neural computation and psychology workshop* (p. 303-311). London: Springer Verlag.

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

Markman, A. B. (1998). *Knowledge representation*. Lawrence Erlbaum Associates.

Masson, M. E. (1986). Comprehension of rapidly presented sentences: The mind is quicker than the eye. *Journal of MEmory and Language*(25), 588–604.

McDonald, S. (2000). *Environmental determinants of lexical processing effort.* Unpublished doctoral dissertation, University of Edinburgh.

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*, *126*, 99–130.

Metcalfe, E. J. (1990). A compositive holographic associative recall model. *Psychological Review*, *88*, 627–661.

Mitchell, J., & Lapata, M. (2008). Vector-based models of semantic composition. In *Proceedings of acl-08: Hlt* (pp. 236–244). Columbus, Ohio.

Montague, R. (1974). English as a formal language. In R. Montague (Ed.), *Formal philosophy.* New Haven, CT: Yale University Press.

Morris, R. K. (1994). Lexical and message-level sentence context effects on fixation times in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 92–103.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*(289–316).

Nelson, D. L., McEvoy, C. L., & Schreiber, T. (1999). *The university of South Florida word association norms.* (`http://www.usf.edu/Freeassociation`)

Neville, H., Nichol, J. L., Barss, A., Forster, K. I., & Garrett, M. F. (1991). Syntactically based sentence processing classes: evidence from event-related brain potentials. *Journal of Cognitive Neuroscience*, *3*, 151–165.

Nosofsky, R. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 104–114.

Nosofsky, R. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.

Nunberg, G., Sag, I., & Wasow, T. (1994). Idioms. *Language*, *70*, 491–538.

O'Seaghdha, P. G. (1989). The dependence of lexical relatedness effects on syntactic connectedness. *Journal of Experiment Psychology: Learning, Memory and Cognition*, *15*, 73–87.

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

Padó, S., & Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, *33*(2), 161–199.

Partee, B. (1995). Lexical semantics and compositionality. In L. Gleitman & M. Liberman (Eds.), *Invitation to cognitive science part i: Language* (pp. 311–360). Cambridge, MA: MIT Press.

Partee, B. (2004). *Compositionality in formal semantics*. Oxford, UK: Blackwell Publishing.

Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). WordNet::similarity – measuring the relatedness of concepts. In *HLT-NAACL 2004: Demonstration Papers* (pp. 38–41). Boston, MA.

Pinker, S. (1994). *The language instinct: How the mind creates language*. New York: HarperCollins.

Plate, T. A. (1991). Holographic reduced representations: Convolution algebra for compositional distributed representations. In J. Mylopoulos & R. Reiter (Eds.), *Proceedings of the 12th international joint*

conference on artificial intelligence, sydney, australia, august 1991 (pp. 30–35). San Mateo, CA: Morgan Kaufmann.

Plate, T. A. (1995, May). Holographic reduced representations [Paper]. *IEEE Transactions on Neural Networks*, *6*(3), 623–641.

Plate, T. A. (2000). Analogy retrieval and processing with distributed vector representations. *Expert Systems: The Journal of Knowledge Engineering*, *17*(1), 29–40.

Pollack, J. B. (1990). Recursive distributed representations. *Artifical Intelligence*, *46*(1–2), 77–105.

Pullum, G. K., & Scholz, B. C. (2007). Systematicity and natural language syntax. *Croatian Journal of Philosophy*, *7*(21), 375–402.

Pynte, J., New, B., & Kennedy, A. (2008). A multiple regression analysis of syntactic and semantic influences in reading normal text. *Journal of Eye Movement Research*, *2(1)*(4), 1–11.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, *88*, 93–134.

Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 95-130.

Resnik, P., & Diab, M. (2000). Measuring verb similarity. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (p. 399-404). Mahwah, NJ: Lawrence Erlbaum Associates.

Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of Verbal Learning and Verbal Behaviour*, *14*, 665–681.

Ross, B. H. (1984). Remindings and their effects in learning a cognitive skill. *Cognitive Psychology*, *16*(371–416).

Ross, B. H. (1987). This is like that: the use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 629–639.

Ross, B. H. (1989a). Distinguishing types of superficial similarities: Different effects on the access and use of earlier problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 456–468.

Ross, B. H. (1989b). Distinguishing types of superficial similarities: Different effects on the access and use of earlier problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(3), 456–468.

Rubenstein, H., & Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, *8*(10), 627–633.

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1). Cambridge MA: MIT Press.

Sahlgren, M., Host, A., & Kanerva, P. (2008). Permutations as a means to encode order in word space. In *Proceedings of the 30th annual meeting of the cognitive science society* (pp. 1300–1305). Mahwah, New Jersey: Erlbaum.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620.

Simpson, G. B., Peterson, R. R., Casteel, M. A., & Brugges, C. (1989). Lexical and context effects in word recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *15*(88–97).

Sloman, S. A., & Rips, L. J. (1998). Similarity as an explanatory construct. *Cognition*, *65*, 87–101.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts*. Cambridge, MA: Harvard University Press.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, *46*, 159–216.

Spenader, J., & Blutner, R. (2007). Compositionality and systematicity. In G. Bouma, I. Krmer, & J. Zwarts (Eds.), *Cognitive foundations of interpretation* (pp. 163–174). Amsterdam: KNAW publications.

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, *29*, 41–78.

Till, R. E., Mross, E. F., & Kintsch, W. (1988). Time course of priming for associate and inference words in discourse context. *Memory and Cognition*, *16*, 283–299.

Wallach, H. M. (2002). *Structured topic models for language*. Unpublished doctoral dissertation, University of Cambridge.

Weeds, J. (2003). *Measures and applications of lexical distributional similarity*. Unpublished doctoral dissertation, University of Sussex, Brighton.

Weiss, S. M., & Kulikowski, C. A. (1991). *Computer systems that learn: Classification and prediction methods from statistics, neural nets, machine learning, and expert systems*. San Mateo, CA: Morgan Kaufmann.

West, R. F., & Stanovich, K. E. (1986). Robust effects of syntactic structure on visual word processing. *Journal of Memory and Cognition*, *14*, 104–112.

Widdows, D. (2008). Semantic vector products: Some initial investigations. In *Proceedings of the second symposium on quantum interaction (qi-2008).* Oxford, UK: College Publications.

Appendix A: Materials

Our experimental stimuli for adjective-noun, noun-noun, and verb-object combinations are shown in Tables 8–10, respectively.

Table 8: Materials for eliciting similarity judgments on adjective-noun combinations.

High
American country–European state, industrial area–whole country, vast amount–large quantity, new body–whole system, small house–little room, early evening–previous day, special circumstance–particular case, black hair–dark eye, new information–further evidence, economic development–rural community, economic problem–practical difficulty, new law–public building, general principle–basic rule, central authority–local office, older man–elderly woman, high price–low cost, different kind–various form, old person–elderly lady, better job–good place, new life–early age, certain circumstance–economic condition, earlier work–early stage, federal assembly–national government, effective way–efficient use, social activity–political action, similar result–good effect, major issue–social event, different part–northern region, important part–significant role, new situation–present position, right hand–left arm, general level–high point, large number–great majority, long period–short time, hot weather–cold air, modern language–new technology

Medium
new life–modern language, good place–high point, social activity–economic condition, different part–various form, better job–good effect, old person–right hand, local office–new technology, high price–short time, social event–low cost, early stage–long period, efficient use–significant role, national government–cold air, large number–vast amount, economic problem–new situation, new information–general level, small house–important part, European state–present position, political action–economic development, large quantity–great majority, dark eye–left arm, northern region–industrial area, little room–similar result, major issue–American country, hot weather–further evidence, new law–basic rule, certain circumstance–particular case, older man–new body, previous day–early age, earlier work–early evening, public building–central authority, elderly woman–black hair, different kind–whole system, effective way–practical difficulty, whole country–general principle, rural community–federal assembly, special circumstance–elderly lady

Low
new situation–different kind, effective way–important part, general level–federal assembly, central authority–political action, major issue–earlier work, older man–great majority, large number–certain circumstance, general principle–present position, similar result–basic rule, northern region–early age, left arm–elderly woman, hot weather–elderly lady, new law–modern language, previous day–long period, whole country–different part, social activity–whole system, new technology–public building, high point–particular case, social event–special circumstance, new body–significant role, early evening–good effect, black hair–right hand, practical difficulty–cold air, short time–rural community, new life–economic development, small house–old person, local office–industrial area, national government–new information, efficient use–little room, various form–European state, better job–economic problem, economic condition–American country, early stage–dark eye, large quantity–good place, vast amount–high price, further evidence–low cost

Table 9: Materials for eliciting similarity judgments on noun-noun combinations.

High
development plan–action programme, telephone number–phone call, marketing director–assistant manager, support group–computer system, training programme–education course, training college–education officer, planning committee–education authority, oil industry–computer company, health service–community care, wage increase–tax rate, environment secretary–defence minister, office worker–health minister, tv set–television programme, party official–government leader, state control–government intervention, tax charge–interest rate, news agency–intelligence service, service department–personnel manager, research work–development project, company director–assistant secretary, labour cost–capital market, league match–football club, world economy–market leader, state benefit–housing department, town council–city centre, party leader–opposition member, management structure–datum system, kitchen door–bedroom window, committee meeting–board member, town hall–county council, research contract–future development, railway station–bus company, care plan–business unit, study group–management skill, tax credit–family allowance, security policy–housing benefit

Medium
state control–town council, party official–opposition member, intelligence service–bus company, state benefit–county council, interest rate–business unit, government intervention–party leader, research work–city centre, capital market–future development, football club–town hall, market leader–board member, tv set–bedroom window, labour cost–housing benefit, care plan–action programme, management structure–computer system, datum system–support group, study group–computer company, research contract–training programme, security policy–defence minister, family allowance–tax rate, tax credit–wage increase, management skill–planning committee, committee meeting–phone call, railway station–oil industry, kitchen door–office worker, education authority–service department, development plan–television programme, community care–tax charge, assistant manager–company director, marketing director–personnel manager, health service–assistant secretary, education officer–development project, education course–housing department, health minister–government leader, telephone number–league match, environment secretary–news agency, training college–world economy

Low
development project–care plan, television programme–research contract, government leader–security policy, tax charge–datum system, news agency–study group, world economy–management structure, assistant secretary–committee meeting, company director–tax credit, league match–family allowance, service department–railway station, housing department–kitchen door, personnel manager–management skill, bus company–health service, city centre–community care, business unit–development plan, town hall–education course, future development–telephone number, party leader–environment secretary, town council–education authority, board member–assistant manager, bedroom window–education officer, county council–marketing director, opposition member–health minister, housing benefit–training college, action programme–tv set, support group–interest rate, tax rate–market leader, training programme–research work, defence minister–government intervention, office worker–party official, computer company–intelligence service, computer system–state control, oil industry–capital market, planning committee–football club, phone call–state benefit, wage increase–labour cost

Table 10: Materials for eliciting similarity judgments on verb-object combinations.

High
produce effect–achieve result, require attention–need treatment, present problem–face difficulty, leave company–join party, satisfy demand–meet requirement, use power–exercise influence, shut door–close eye, sell property–buy land, reduce amount–increase number, send message–receive letter, suffer loss–cause injury, use test–pass time, write book–read word, start work–begin career, reach level–achieve end, stress importance–emphasise need, use method–develop technique, hold meeting–attend conference, use knowledge–acquire skill, win match–play game, like people–ask man, follow road–cross line, help people–encourage child, pose problem–address question, raise head–lift hand, pay price–cut cost, leave house–buy home, wave hand–stretch arm, discuss issue–consider matter, provide help–offer support, win battle–fight war, remember name–hear word, set example–provide system, provide datum–collect information, pour tea–drink water, share interest–express view

Medium
write book–hear word, address question–raise head, read word–remember name, follow road–set example, use method–drink water, hold meeting–lift hand, win match–fight war, play game–win battle, start work–wave hand, achieve end–express view, develop technique–provide help, attend conference–share interest, provide system–use power, cut cost–reduce amount, buy home–sell property, consider matter–produce effect, leave house–buy land, pay price–require attention, collect information–receive letter, offer support–need treatment, discuss issue–present problem, stretch arm–close eye, pour tea–join party, provide datum–shut door, face difficulty–pose problem, achieve result–reach level, exercise influence–use knowledge, satisfy demand–emphasise need, send message–ask man, use test–acquire skill, meet requirement–help people, leave company–encourage child, pass time–cross line, suffer loss–begin career, increase number–like people, cause injury–stress importance

Low
use knowledge–provide system, pose problem–consider matter, encourage child–leave house, reach level–provide datum, ask man–stretch arm, acquire skill–buy home, stress importance–cut cost, begin career–pay price, cross line–offer support, help people–discuss issue, like people–collect information, emphasise need–pour tea, drink water–use test, remember name–pass time, share interest–exercise influence, hear word–send message, wave hand–leave company, fight war–increase number, provide help–satisfy demand, raise head–cause injury, lift hand–achieve result, set example–face difficulty, express view–suffer loss, win battle–meet requirement, buy land–write book, receive letter–read word, produce effect–start work, present problem–address question, use power–develop technique, sell property–hold meeting, shut door–follow road, join party–play game, close eye–achieve end, reduce amount–win match, need treatment–use method, require attention–attend conference

Appendix B: Experimental Instructions

In this experiment you will be shown a pair of noun phrases. Your task is to judge how similar the two phrases are. You will make this judgement by choosing a rating from 1 (not very similar) to 7 (very similar). The focus is on the similarity of the concepts named by the phrases, not any association between the two phrases.

For example, if you were asked to make the following comparison:

(1) a. professional advice
    b. expert opinion

you would give this a high similarity rating (e.g. 6 or 7). Both these phrases concern guidance or instruction from a knowledgeable person and so have highly similar meanings. On the other hand, if you were given the following comparison:

(2) a. social worker
    b. wide range

you would probably choose a low similarity rating (e.g. 1 or 2), since one is an occupation and the other is a magnitude. Likewise, for this comparison:

(3) a. increasing taxation
    b. public protest

you would also choose a low similarity rating (e.g. 1 or 2), since they are different things, even though they might be associated, in that the first could lead to the second. Of course, associated phrases may also be similar.

Sometimes the two phrases will have meanings that are moderately different though still have much in common. For instance, in this comparison:

(4) a. human behaviour
    b. social activity

you would choose a middling rating (e.g., 3, 4 or 5) if you felt that the meanings of the two phrases were reasonably different but also had some similarities. For instance both involve the interactions of people, although the two phrases also invoke other distinct concepts.

There are no 'correct' answers, so whatever choice seems appropriate to you is a valid response. Simply try to rate how similar the meanings of the two phrases are. Base your judgment on your first impression of what each phrase means. The whole experiment should take only about 10 minutes.

Remember:

- Rate the similarity of the phrases not their association.
- Base your judgment on your first impression of what each phrase means.
- There are no correct answers.

At the start of the experiment you will be given a few examples to practice on.

Appendix C: Simple Vector and Tensor Algebra

Formally, vectors are defined as objects within a vector space, which is itself defined in terms of the abstract relations and properties of the objects it contains. Informally, we usually think of a vector as something having magnitude and direction, such as an arrow in space. From the point of view of computation, we can always represent vectors concretely in terms of some particular basis, that is as a set of co-ordinates. Thus, a vector $\mathbf{v}$ is represented by its components v_i in that basis.

An important relation between vectors is the dot product, defined as the sum of the products of the components:

$$\mathbf{u} \cdot \mathbf{v} = \sum_i u_i v_i \tag{30}$$

where the index i ranges over all the components (i.e., the dimensions of the space). This allows us to define the length of vectors:

$$|\mathbf{v}| = \sqrt{\mathbf{v} \cdot \mathbf{v}} \tag{31}$$

The dot product also has a useful relation to the angle, θ, between the two vectors:

$$\mathbf{u} \cdot \mathbf{v} = |\mathbf{u}||\mathbf{v}| cos(\theta) \tag{32}$$

This implies that the dot product of any two orthogonal vectors ($\theta = 90^\circ$) is zero. Equations (30)–(32) allow us to calculate the cosine of the angle as:

$$cos(\theta) = \frac{\sum_i u_i v_i}{\sqrt{\sum_i u_i u_i}\sqrt{\sum_i v_i v_i}} \tag{33}$$

The two most basic transformations which operate on vectors are addition and multiplication by a scalar:

$$\mathbf{p} = \mathbf{u} + \mathbf{v} \tag{34}$$

$$p_i = u_i + v_i \tag{35}$$

$$\mathbf{q} = s\mathbf{v} \tag{36}$$

$$q_i = s v_i \tag{37}$$

Any vector $\mathbf{v}$, can be expressed in terms of a component, $\mathbf{v}_\parallel$, parallel to and a component, $\mathbf{v}_\perp$, orthogonal to a second vector, $\mathbf{u}$.

$$\mathbf{v} = \mathbf{v}_\parallel + \mathbf{v}_\perp \tag{38}$$

Taking the dot product of $\mathbf{u}$ on both sides of this equation yields:

$$\mathbf{v} \cdot \mathbf{u} = |\mathbf{v}_\parallel||\mathbf{u}| Cos(0^\circ) + |\mathbf{v}_\perp||\mathbf{u}| Cos(90^\circ) \tag{39}$$

Since $Cos(0^\circ) = 1$ and $Cos(90^\circ) = 0$, this implies:

$$|\mathbf{v}_\parallel| = \frac{\mathbf{v} \cdot \mathbf{u}}{|\mathbf{u}|} \tag{40}$$

Thus, we can construct $\mathbf{v}_\parallel$ by normalizing $\mathbf{u}$ to give a unit vector which points in the right direction, and then multiplying by the right magnitude $|\mathbf{v}_\parallel|$.

$$\mathbf{v}_\parallel = \frac{\mathbf{v} \cdot \mathbf{u}}{|\mathbf{u}|} \frac{\mathbf{u}}{|\mathbf{u}|} = \frac{\mathbf{v} \cdot \mathbf{u}}{\mathbf{u} \cdot \mathbf{u}} \mathbf{u} \tag{41}$$

The orthogonal component, $\mathbf{v}_\perp$, can then be calculated from the fact that the two components must combine to give $\mathbf{v}$.

$$\mathbf{v}_\perp = \mathbf{v} - \frac{\mathbf{v}\cdot\mathbf{u}}{\mathbf{u}\cdot\mathbf{u}}\mathbf{u} \tag{42}$$

Another important set of transformations of vectors are the linear transformations induced by matrices.

$$\mathbf{w} = \mathbf{M}\mathbf{v} \tag{43}$$

$$w_i = \sum_j M_{ij} v_j \tag{44}$$

A matrix, $\mathbf{M}$, is represented by an array of values, M_{ij}, indexed by a pair of indices, i and j. The element M_{ij} can be thought of as determining how much the jth component of the original vector, $\mathbf{v}$, contributes to the ith component of the transformed vector, $\mathbf{w}$.

Alternatively, we can think of $\mathbf{M}$ in terms of its polar decomposition into a matrix, $\mathbf{D}$, of dilations, and a matrix, $\mathbf{R}$, of rotations.

$$\mathbf{M} = \mathbf{D}\mathbf{R} \tag{45}$$

The action of $\mathbf{D}$ on a vector is to stretch it by various amounts in various directions, and the action of $\mathbf{R}$ is to rotate it around various axes, without changing its length. The directions in which a matrix dilates vectors without changing their direction are known as eigenvectors, and the amounts by which they are stretched are known as eigenvalues.

The tensor product, $\otimes$, takes a pair of vectors and combines them to form a higher dimensional vector.

$$\mathbf{t} = \mathbf{u}\otimes\mathbf{v} \tag{46}$$

In particular, if $\mathbf{u}$ and $\mathbf{v}$ have dimension n, then $\mathbf{t}$ has dimension n^2. The components of $\mathbf{t}$ are all the pairwise products of the components of $\mathbf{u}$ and $\mathbf{v}$.

$$t_{ij} = u_i v_j \tag{47}$$

It is possible to project such a product of vectors down onto a vector of the same dimension using a rank 3 tensor.

$$\mathbf{r} = \mathbf{C}\mathbf{u}\mathbf{v} \tag{48}$$

Again this is a linear transformation. The tensor $\mathbf{C}$ has three indices, corresponding to the three vectors $\mathbf{r}$, $\mathbf{u}$ and $\mathbf{v}$.

$$r_i = \sum_{jk} C_{ijk} u_k v_j \tag{49}$$

A simple example of such a rank 3 tensor would be one in which $C_{ijk} = 1$ when $i = j = k$ and 0 otherwise, which yields:

$$r_i = u_i v_i \tag{50}$$

which can be also written as:

$$\mathbf{r} = \mathbf{u}\odot\mathbf{v} \tag{51}$$

A more complex example is the tensor, $\mathbf{C}$, with components $C_{ijk} = 1$ when $k = (i-j) mod_n$ and 0 otherwise.

$$r_i = \sum_j u_j v_{(i-j) mod_n} \tag{52}$$

also written as:

$$\mathbf{r} = \mathbf{u} \circledast \mathbf{v} \tag{53}$$

Multiplying a single vector, **u**, by a rank 3 tensor, **C**, produces a matrix, **U**.

$$\mathbf{U} = \mathbf{Cu} \tag{54}$$

$$U_{ij} = \sum_k C_{ijk} u_k \tag{55}$$

Multiplication of **v** by this matrix, **U**, then results in the same vector, **r**, produced by the product **Cuv**:

$$\mathbf{Uv} = \mathbf{Cuv} = \mathbf{r} \tag{56}$$

$$\sum_j U_{ij} v_j = \sum_j \sum_k C_{ijk} u_k v_j = r_i \tag{57}$$

It is also possible to define higher order tensors, such as a rank 4 tensor which acts on three vectors:

$$\mathbf{x} = \mathbf{Duvy} \tag{58}$$

$$x_i = \sum_{jkl} D_{ijkl} u_l v_k y_j \tag{59}$$