Language and Linguistics Compass 6/10 (2012): 635–653, 10.1002/lnc3.362

Vector Space Models of Word Meaning and Phrase Meaning: A Survey

Katrin Erk*
Department of Linguistics, The University of Texas at Austin

Abstract

Distributional models represent a word through the contexts in which it has been observed. They can be used to predict similarity in meaning, based on the distributional hypothesis, which states that two words that occur in similar contexts tend to have similar meanings. Distributional approaches are often implemented in vector space models. They represent a word as a point in high-dimensional space, where each dimension stands for a context item, and a word's coordinates represent its context counts. Occurrence in similar contexts then means proximity in space. In this survey we look at the use of vector space models to describe the meaning of words and phrases: the phenomena that vector space models address, and the techniques that they use to do so. Many word meaning phenomena can be described in terms of semantic similarity: synonymy, priming, categorization, and the typicality of a predicate's arguments. But vector space models can do more than just predict semantic similarity. They are a very flexible tool, because they can make use of all of linear algebra, with all its data structures and operations. The dimensions of a vector space can stand for many things: context words, or non-linguistic context like images, or properties of a concept. And vector space models can use matrices or higher-order arrays instead of vectors for representing more complex relationships. Polysemy is a tough problem for distributional approaches, as a representation that is learned from all of a word's contexts will conflate the different senses of the word. It can be addressed, using either clustering or vector combination techniques. Finally, we look at vector space models for phrases, which are usually constructed by combining word vectors. Vector space models for phrases can predict phrase similarity, and some argue that they can form the basis for a general-purpose representation framework for natural language semantics.

1. Introduction

We can often guess what a word means from the contexts in which it is used. Thus, we can represent meaning as *distribution*, as observed contexts. The origin of this notion can be traced back to Wittgenstein (1997), who wrote that ''the meaning of a word is its use in the language''. The idea of meaning as distribution is put into practice in *distributional models*, which represent the meaning of a word through the contexts in which it has been observed in a corpus. The definitions of what a context is differ widely; in the simplest case, a context is just a bag of words that have occurred in the vicinity of the target word. Distributional models predict similarity in meaning based on similarity of observed contexts, based on the *distributional hypothesis* (Harris 1954; Firth 1957): If two words tend to occur in similar contexts, we can assume that they are similar in meaning. Distributional models are typically implemented through *vector space models*,[1] where the representation for a word is a point in a high-dimensional space. The dimensions stand for context items (for example, co-occurring words), and the coordinates depend on the co-occurrence counts. Contextual similarity then becomes proximity in space. Distributional models can be learned from a corpus in an unsupervised fashion.

This survey looks at the use of vector space models for representing word and phrase meaning: the phenomena that can be modeled, and the techniques that are being used. For a more fundamental introduction to the computation and use vector space models, see the recent survey by Turney and Pantel (2010). In this paper, we discuss models for representing meaning at three levels of granularity: words, individual word occurrences, and phrases. At the level of words, there are many word meaning phenomena that can be described in terms of semantic similarity, including synonymy, priming, categorization, and the typicality of a predicate's arguments. But vector space models can do more than just predict semantic similarity. They can make use of all the data structures and operations of linear algebra. Also, the dimensions of a vector space can stand for more than just context words, for example non-linguistic context like images, or properties of a concept. And vector space models can use matrices or higher-order arrays instead of vectors for representing more complex relationships. At the level of individual word occurrences, vector space models can address the problem of polysemy. Polysemy is a tough problem for distributional approaches, as a representation that is learned from all of a word's contexts will conflate the different senses of the word. So vectors for individual word occurrences are usually constructed using either clustering, vector combination, or a language model. Word occurrence vectors have one important advantage: They can handle polysemy without needing to refer to dictionaries. At the level of phrases, we look at models that can predict phrase similarity. Again, the construction of such models is not straightforward. Many phrases do not occur with sufficient frequency in a corpus to be represented through their distributional contexts. Instead, a phrase representation is constructed from the vectors for the words that occur in it. Among the phrase-level models are some with a very ambitious goal: the design of a general-purpose sentence semantics on the basis of vector space models.

2. *Constructing Vector Space Models of Distributional Context*

AN EXAMPLE

Distributional models represent meaning through observed contexts. Figure 1 shows a simple example for the target word ''bathtub''. The toy corpus on the left consists of a few sentences from the British National Corpus. In the middle are co-occurrence counts derived from the toy corpus (showing counts for only some of the context words). For this example, we have counted lemmatized context words in the full sentence in which the target ''bathtub'' occurs. Distributional models are commonly implemented in vector space models that represent a target word – here, ''bathtub'' – as a point in high-dimensional space. The dimensions correspond to the context items, and in the simplest case, the coordinates are the co-occurrence counts. Figure 1 (right) shows parts of the vector for ''bathtub'' derived from the toy corpus.

PARAMETERS

Vector space models are characterized by many parameters. Lowe (2001) sums up the most important ones by defining a vector space as a tuple $\langle A,B,S,M\rangle$. B is the set of basis elements or dimensions, for example context words, or whole documents. S is a similarity measure between pairs of vectors. It predicts semantic similarity between words as proximity between their vectors. One widely used measure is the cosine of the angle between the two vectors, illustrated in Figure 2. This measure will consider two words similar if

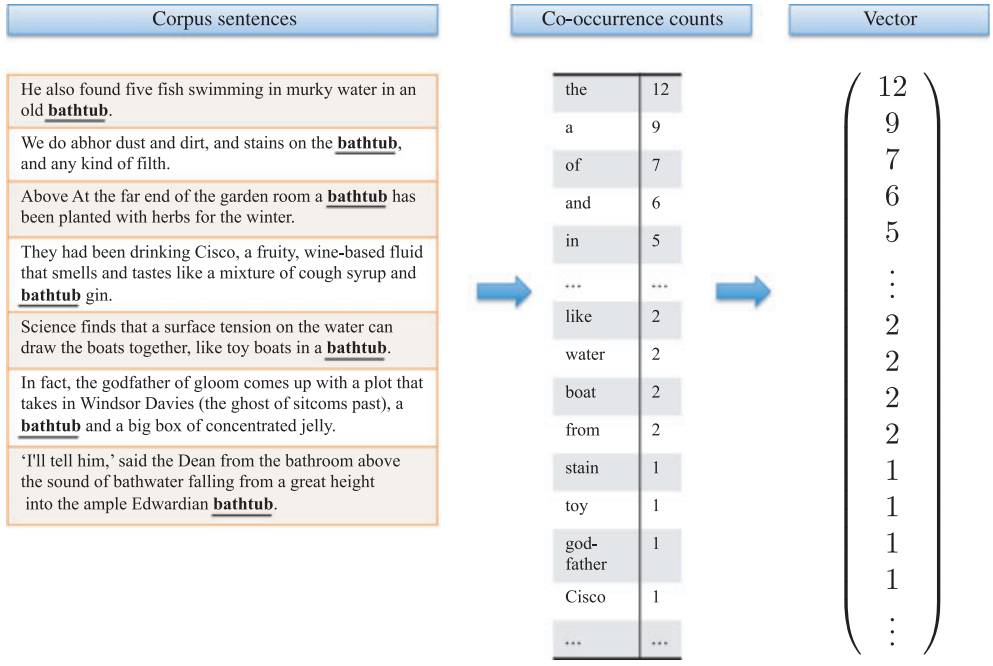


Fig 1. Creating a simple vector space representation for bathtub: A toy corpus of sample sentences from the British National Corpus (left), context word counts (middle), and the corresponding vector (right).

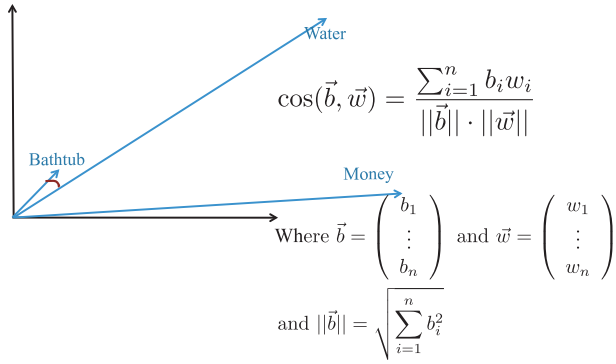


Fig 2. Computing similarity between ''bathtub'' and ''water'' as the cosine of the angle between the vectors $\vec{b}$ for ''bathtub'' and $\vec{w}$ for ''water''. The word ''water'' is much more frequent overall than ''bathtub'', but they will still be judged similar. The word ''money'' has similar frequency as ''water'', but we would like it to be predicted to be less similar to ''water'' than ''bathtub''. This is a toy 2-dimensional example, due to the limits of visualization.

they have context words in similar proportions, irrespective of their overall frequency. A is a function that changes raw co-occurrence counts to association weights in order to reduce noise, for example pointwise mutual information. Finally, M is a transformation of the whole vector space, typically dimensionality reduction. As an example of a popular vector space model, Latent Semantic Analysis (LSA, Landauer and Dumais 1997), uses documents as basis vectors B and uses tf/idf as the association function A. The whole space is transformed using singular value decomposition as the mapping M, yielding latent semantic classes as dimensions. The similarity measure S that LSA uses is cosine.

While most approaches are based on vectors, some use matrices or even higher order tensors. Tensors are generalizations of vectors and matrices: A first-order tensor is a vector, a second-order tensor is a matrix, a third-order tensor is a three-dimensional array, and so on.

BAYESIAN APPROACHES

Recently, Bayesian approaches that use distributional information have become popular, in particular Latent Dirichlet Allocation topic models (Blei et al. 2003). They learn *topics* as distributions over words, and describe a document as a mixture of topics. Topics can be interpreted by looking at high-probability words. For example, a topic in which words like ''new'', ''film'', ''music'', ''movie'' have high probability could be characterized as ''Arts''. Under a topic model, a word can be characterized as a point in a vector space in which the dimensions are topics, and a word's coordinates are its probabilities under all the topics (e.g., Feng and Lapata 2010; Dinu and Lapata 2010). Topic models are similar to dimensionality reduction: Like LSA and pLSA (Deerwester et al. 1990; Hofmann 1999), they relate both documents and word to latent classes, in this case the topics.

APPLICATIONS

The idea of modeling semantic similarity as similarity of context has been extremely successful in computational linguistics and has been applied in a wide variety of applications, including information retrieval (Deerwester et al. 1990; Manning et al. 2008), word sense induction (Schütze 1998), ontology learning (Hindle 1990; Lin 1998; Ravichandran et al. 2005; Gorman and Curran 2006; Snow et al. 2006), determining predominant word sense (McCarthy et al. 2004, Padó and Lapata 2007), predicting similarity of semantic relations (Turney 2006), and learning inference rules (Lin and Pantel 2001). A recent survey article (Turney and Pantel 2010) discusses these and other applications in greater detail.

3. Modeling Word Meaning

WHAT IS WORD SIMILARITY?

The central property of distributional models is that they predict the semantic similarity of words based on their contextual similarity. But what is word similarity? Figure 3 (left) shows a few datapoints from the Rubenstein and Goodenough (1965) dataset, for which human subjects have rated the similarity of word pairs on a scale. Some of the highly rated pairs are synonyms, like ''midday'' and ''noon''. But similar words can also be hyponym and hypernym, such as ''fruit'' and ''food'', or sister terms in some hierarchy, such as ''king'' and ''rook'', or they may just be connected to a joint scenario, such as ''doctor'' and ''hospital'' '(Budanitsky and Hirst 2006).

MODELING WORD MEANING PHENOMENA WITH DISTRIBUTIONAL SIMILARITY

The fact that there is no single semantic relation that semantic similarity encodes may be an advantage for distributional approaches, as it enables them to model many phenomena through distributional similarity. Besides word meaning judgments (e.g., McDonald and Ramscar 2001), distributional similarity can detect synonyms (Landauer and Dumais

Rubenstein/Goodenough dataset: Human judgments on word similarity

Word 1	Word 2	Average sim
midday	Noon	3.94
grin	Smile	3.46
food	Fruit	2.69
coast	Hill	1.26

TOEFL synonyms sample question

Stem:	levied
choices:	(a) imposed (b) believed (c) requested (d) correlated
Solution:	(a) imposed

Plausibility of role fillers for the agent and patient role of verbs (McRae et al, 1998)

Verb	Noun	Role	Rating
Fire	employer	Agent	6.1
Fire	employer	Patient	2.4
Fire	employee	Agent	1.9
Fire	employee	Patient	6.4

Fig 3. Data modeled by distributional similarity. Left: human judgments on word similarity (5-point scale) (Rubenstein and Goodenough 1965). Middle: sample Test of English as a Foreign Language (TOEFL) synonym question. Right: human plausibility judgments for role fillers of predicates (7-point scale, averaged over participants), from McRae et al. (1998).

1997) in a task illustrated in Figure 3 (middle). A distributional model will select the option most similar to the probe as the synonym. Another phenomenon that can be modeled is *priming* in lexical decision tasks (Lund et al. 1995): Human subjects are faster in making a decision about a word – for example, decide whether it is a word of the English language – if they have just encountered a semantically related word. Distributional models predict words to be the more primed the more similar they are. Burgess and Lund (1997) simulate categorization based on semantic similarity. Among words for animals, cities, geographic locations, and body parts, words from the same category tended to be more similar than words from different categories. And Padó et al. (2007) use semantic similarity to predict selectional preferences. Selectional preferences describe the typicality of arguments for a given verb, as illustrated in Figure 3 (right): ''employer'' is a more typical agent for the verb ''fire'' than ''employee''. Padó et al. compile a list of seen arguments for each role (like the agent role of ''fire''), then judge the typicality of a new argument based on its similarity to the previously seen ones.

ARE MENTAL CONCEPTS DISTRIBUTIONAL?

Distributional models can describe many word meaning phenomena. Is this because they are built from text corpora that reflect the world knowledge that is also encoded in mental concepts? Or is it because mental concepts themselves are distributional in nature? One fact that points in the latter direction is that human subjects are able to induce the meaning of unknown words from distributional context (McDonald and Ramscar 2001). And especially for abstract concepts like ''truth'' or ''government'', it seems plausible that language interaction would play an important role in learning word meaning (Landauer and Dumais 1997; Andrews et al. 2009). Others argue that a framework that only characterizes symbols through other symbols, without any reference to the world, cannot constitute meaning. Theories of *embodied cognition* propose that concepts are organized in the brain based on their sensory and motor properties, and that simulation of actions and perceptions in the brain plays a central role in cognition (Barsalou 2008). It is also possible that both embodied cognition and relation between symbols play a role (Andrews et al. 2009), or that the two are interdependent and mutually reinforcing because language has evolved to encode embodied information (Louwerse 2007, 2010).

BEYOND BAG-OF-WORDS TEXTUAL CONTEXT

All phenomena that we have discussed so far could be addressed using semantic similarity, where the similarity was predicted by a distributional model trained on bag-of-words textual contexts. Next we turn to models that go beyond this simple standard framework.

VECTORS FOR WORD SEQUENCES

Jones and Mewhort (2007) construct vectors for word n-grams, and then derive the vector for a target word *w* as the superposition (addition) of n-gram vectors for observed contexts of *w*. While the distributional models that we have considered so far treat documents as unordered bags of words, the model by Jones and Mewhort can address word order phenomena, like knowing that ''Hong'' is highly likely to be followed by ''Kong'' but unlikely to be preceded by it.

NON-TEXTUAL CONTEXT

Distributional context need not be solely textual. For example, texts are often illustrated with images. Several recent approaches combine textual context words with *visual words*, which represent an image as a bag of keypoint features (Feng and Lepata 2010; Leong and Mihalcea 2011; Bruni et al. 2011). The resulting vector space models then have mixed textual and visual dimensions. For target words that are easy to visualize, these models do better than text-only models at estimating semantic similarity. The models can also be used for tasks that link images and text, such as automatic image labeling and text illustration.

DIMENSIONS AS PROPERTIES

Given a word such as ''strawberry'', human subjects can generate lists of defining features, called *feature norms* (Vigliocco et al. 2004; McRae et al. 2005). Figure 4 (left) shows the feature norms that were elicited for ''strawberry''. Based on semantic similarity alone, distributional models cannot generate such features with good accuracy (Baroni et al. 2010). But feature norm-like expressions can be extracted from corpora (Almuhareb and Poesio 2004; Devereux et al. 2009; Baroni and Lenci 2010; Baroni et al. 2010; Kremer and Baroni 2010). Most approaches do this by using text pattens or dependency parse patterns. For example, from a sequence ''the red strawberry was…'' we can extract ''red'' as a property for ''strawberry''. The resulting property list can again be used as a vector space representation, and can even serve to make similarity predictions (Baroni et al. 2010). But in addition, such models can make predictions based on coordinates for individual dimensions. For example, they can be used to predict feature norms of a word as those dimensions with high weight. Figure 4 (right) shows an example of such a representation.[2]

COMPOSITE DIMENSIONS

The models by Baroni et al. (2010) and Baroni and Lenci (2010) (the latter shown in Figure 4 right) additionally extend the vector space framework by moving away from atomic dimensions. Their dimensions are pairs of a relation and a context word, for example ''in drown-v''. The representation for all target words can be viewed as one

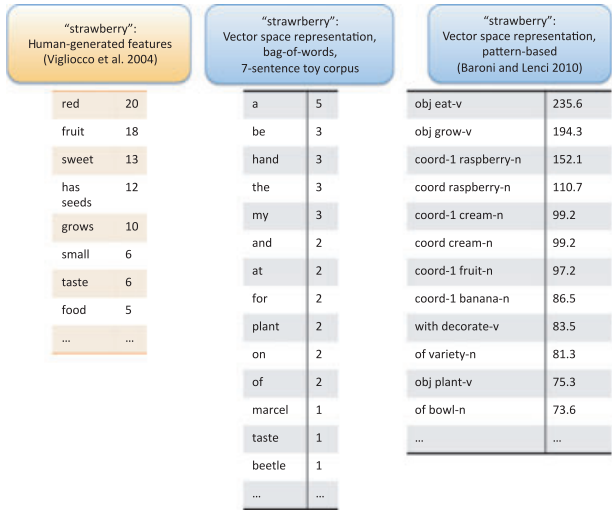


Fig 4. Interpretable versus opaque features for the noun ''strawberry'': Left, human-generated feature norms, from Vigliocco et al. (2004). Numbers indicate the number of subjects that named each feature. Middle, bag-of-words vector with counts from a 7-sentence toy corpus. Right, pattern-based vector, from Baroni and Lenci (2010).

single third-order tensor, a 3-dimensional array in which the cells are addressed by the target word, the relation, and the context word. The model can be adapted to different tasks by restricting it to cells with particular labels. For example, selectional preferences (typical arguments) for the direct object role can be read off if we select all cells with verb targets and the relation ''obj''.[3]

CONCEPTUAL SPACES

The theoretical model of Gärdenfors (2004) marks a more radical departure from bag-of-words distributional context than the models we have seen so far. In his *conceptual spaces*, dimensions stand for *qualities* like the hue, saturation, and brightness of a color. In a space with these three dimensions, each point uniquely characterizes a color. So one fundamental difference from distributional models is that each point can be interpreted as a potential entity or attribute. A *property* is a region in space, for example the region of all points that can be considered ''red''. This, too, is different from most distributional models, which represent words as points rather than regions. Gärdenfors' conceptual spaces are vector spaces, but they are not distributional. Erk (2009) transfers one aspect of Gärdenfors' model to a distributional setting, the representation of words as a region rather than a point. To induce a region for a word, vectors (points in space) are first computed for individual occurrences of the word (see the following section for techniques). The region is then estimated from the locations of these occurrence points in space.

4. *Characterizing the Meaning of Individual Word Occurrences*

Many words have more than one meaning. An example is the noun ''track'', which can be used to mean ''path'', or a piece of music on a sound recording medium, or it can refer to running. Figure 5 shows a toy corpus of a few sample uses of the noun. If a word vector for ''track'' is induced from this toy corpus, it will conflate the different meanings

The green swarded **track** alternated between old military road and the disused railway **track** foundations to Fort Augustus.
Things are hardly helped by new singer Paul Roberts, who delivers almost every **track** in overblown, melodramatic tones that are downright hilarious.
In fact, given his **track** record, even his real name probably wasn't his real name.
That exchange came to mind as I watched Gavin Scott's Horizon, British science — on the wrong **track**?
Reactivity with p46 and p55 was only present in the immune (I, **track** 2) but not preimmune (PI, **track** 3) serum.
'How often do you seriously think about why you are going to the **track**, what workout you are going to do and exactly how it is going to help you achieve peak performance on the **track** just when it counts?
Most of the British network is double **track**, while in Spain, four fifths of the network is single **track**.

Fig 5. Polysemy: Sample occurrences of the noun ''track'' in the British National Corpus.

of the word. In this section we discuss models that address this problem by deriving vector space representations for individual word occurrences. Besides providing a way to handle polysemy in vector spaces, these models have the additional advantage of being able to address word meaning without referring to dictionary senses (Erk 2010). The use of a fixed set of dictionary senses has been drawn into question both in terms of cognitive validity and in terms of adequacy for applications (Kilgarriff 1997; Hanks 2000; Kintsch 2007).

VECTOR ADAPTATION

One group of approaches takes as a starting point the word vector for ''track'', computed from all its corpus occurrences and thus conflating the word's meaning. The vector is then modified to fit a particular sentence context (Landauer and Dumais 1997; Kintsch 2001; Mitchell and Lapata 2008; Erk and Padó 2008; Thater et al. 2009, 2010, 2011; Van de Cruys et al. 2011). The simplest method uses vector addition (illustrated in Figure 6a) to combine the word vectors for the target and the surrounding words. In the first sentence of Figure 5, the vectors for ''green'', ''swarded'', ''alternate'', …, would be added to the vector for ''track''. Component-wise multiplication (Figure 6b), while equally simple, has been found to work better (Mitchell and Lapata 2008), at least when vector combination was restricted to a single neighboring word. Component-wise multiplication weights the target's dimensions by importance to the sentence context; dimensions that are not relevant to both the target and the sentence context are set to zero. Vector addition, on the other hand, pools the information of target and sentence context, thereby potentially increasing noise. Both vector addition and component-wise multiplication ignore syntax; some other approaches give syntax a key role. In the context ''*track*

$$\begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} + \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = \begin{pmatrix} v_1 + w_1 \\ v_2 + w_2 \\ v_3 + w_3 \end{pmatrix} \qquad \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} \odot \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = \begin{pmatrix} v_1 \cdot w_1 \\ v_2 \cdot w_2 \\ v_3 \cdot w_3 \end{pmatrix}$$

(a) addition **(b)** component-wise multiplication

$$\begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} \otimes \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = \begin{pmatrix} v_1 w_1 & v_1 w_2 & v_1 w_3 \\ v_2 w_1 & v_2 w_2 & v_2 w_3 \\ v_3 w_1 & v_3 w_2 & v_3 w_3 \end{pmatrix}$$

(c) outer product (tensor product)

Fig 6. Vector combination (illustrated for two 3-dimensional vectors). (a) addition (b) component-wise multiplication (c) outer product (tensor product).

alternated'' (first sentence of Figure 5), Erk and Padó (2008) would adapt the vector of ''track'' to be more similar to typical subjects of ''alternate''. Thater et al. (2009, 2010, 2011) represent the target as a vector whose dimensions correspond to typical arguments. In a particular sentence context, for example ''swarded *track*'', the vector is narrowed down using the actual arguments, here ''swarded''.

CLUSTERING

Another approach to the polysemy problem is to use clustering (Dinu and Lapata 2010; Erk and Padó 2010; Reisinger and Mooney 2010). Each individual occurrence of ''track'' in the corpus matches one of the word's meanings. If we cluster the occurrences, we can automatically learn the meanings (Schütze 1998; Reisinger and Mooney 2010). An individual occurrence can be represented using the methods above, or simply as a vector with counts for the words in only that sentence. For the first sentence in Figure 5, this vector would have a count of 2 for ''the'', 1 for ''green'', and so on. Another related approach computes a topic model over the corpus, resulting in topics that are soft clusters over the words. A word is then represented through its probability under each topic (Dinu and Lapata 2010). The meaning of a single occurrence can then be characterized either through a single cluster, or through a weight distribution over clusters.

LANGUAGE MODELS

A third group of approaches uses a language model to predict other words that can occur in the position occupied by the target (Deschact and Moens 2009; Washtell 2010; Moon and Erk forthcoming). In the expression ''the green swarded *track*'', the words ''the green swarded…'' could also be followed by words like ''battery'', ''parkland'', ''mountains'', ''banks'', or ''Camelot''. These predictions are then used as the representations of ''track'' in this context.

INTERPRETING WORD OCCURRENCE VECTORS THROUGH PARAPHRASES

A vector for a word occurrence cannot be directly interpreted in the way a dictionary sense would be. But it can be located with respect to other vectors, in particular vectors of potential paraphrases. In the first sentence of Figure 5, the (occurrence) vector for ''track'' should be quite close to the vector for ''path'', and not so close to the vectors for ''racecourse'' or ''piece'' (of music). Word occurrence vectors are typically evaluated on a paraphrasing task, mostly using the Lexical Substitution dataset of McCarthy and

Sentence	Substitutes
Some payments occurred " after the traffickers had been indicted by federal law enforcement agencies on drug charges, in others while traffickers were under active investigation by these same agencies. (#1812)	accusation: 2 allegation: 2 offence: 1 indictment: 1
We study the methods and concepts that each writer uses to defend the cogency of legal, deliberative, or more generally political prudence against explicit or implicit charges that practical thinking is merely a knack or form of cleverness. (#1813)	allegation: 3 accusation: 2 criticism: 1

Fig 7. Two sample sentences from the Lexical Substitution data (McCarthy and Navigli 2009). Numbers indicate the number of annotators that proposed each substitute.

Navigli (2009), of which Figure 7 shows an example. Absolute performance figures are low, which indicates that the task is difficult. But it is too early to tell whether it is more or less difficult than word sense disambiguation using dictionary senses. There is currently little paraphrasing data available for evaluation. Also, current occurrence vector models are still relatively simple compared to word sense disambiguation models. With the exception of Van de Cruys et al. (2011), they all use only one single information source: either just bag-of-words context, or just syntactic neighbors, or just n-gram context. This makes them hard to compare to word sense disambiguation models.

5. Vector Space Models for Phrase Meaning

In this section, we move from word models to phrase models. As we will see, there is a notion of phrase similarity that parallels word similarity. Phrase similarity, like word similarity, can be predicted using vector space models. But as sentences grow more complex, the vector space models to represent them must put more emphasis on the encoding of sentence structure. In the second half of this section, we discuss approaches whose eventual goal is to provide a general-purpose sentence semantics on the basis of vector space models.

PHRASE SIMILARITY

Can human subjects give consistent judgments on phrase similarity, like they do for word similarity? Mitchell and Lapata (2010) elicited similarity judgments for three different types of two-word phrases: adjective-noun, noun-noun, and verb-object pairs, illustrated in Figure 8. Similar to the situation with word similarity, there is no single semantic relation that underlies phrase similarity. Some of the high-scoring pairs in Figure 8 are paraphrases, like "vast amount / large quantity". But "old person" is a hypernym of "elderly lady". And the phrases "share interest /express view" are related to a common scenario, but their exact relation is hard to characterize. (See also Washtell (2011) for phrase similarity judgments for longer phrases.)

MODELING PHRASE SIMILARITY

In principle, phrases could be given distributional representations based on the contexts in which they appear in a corpus. But this is usually not done because the counts would

Adjective/noun pairs

Phrase 1	Phrase 2	Average sim
vast amount	large quantity	6.72
old person	elderly lady	5.72
small house	little room	3.83
social event	low cost	1.56

Noun/noun pairs

Phrase 1	Phrase 2	Average sim
state control	government intervention	5.83
health minister	government leader	4.94
business unit	development plant	3.72
tv set	bedroom window	1.61

Verbs and objects

Phrase 1	Phrase 2	Average sim
reduce amount	cut cost	6.56
share interest	express view	5.72
present problem	face difficulty	3.56
encourage child	leave company	128

Fig 8. Sample items from the phrase similarity dataset of Mitchell and Lapata (2010). 7-point scale, numbers shown are averaged over participants.

be too sparse for most phrases. Instead, the representation for a phrase is composed out of representations for its parts. Mitchell and Lapata (2008) focus on two-word phrases and formulate a general framework for vector composition for that case as

$$\vec{p} = f(\vec{u}, \vec{v}, R, K)$$

The vector for the phrase $p = uv$ is a function of the vectors for u and v, the syntactic relation R between the two words, and background knowledge K. To model similarity between two-word phrases, Mitchell and Lapata test several composition methods, among them vector addition and component-wise multiplication,[4] as well as tensor product (illustrated in Figure 6c), ignoring parameters R and K for the time being. They again find good performance for component-wise multiplication. This method has become a standard baseline for other approaches to modeling two-word phrases, many of which focus specifically on modifier/noun pairs (Baroni and Zamparelli 2010; Guevara 2010, 2011; Zanzotto et al. 2010; Reddy et al. 2011; Hartung and Frank 2011; Washtell 2011). Washtell (2011) and Reddy et al. (2011) take the phrase-specific meaning of words into account and compose phrase representations out of word occurrence vectors. While most approaches are unsupervised, Guevara (2010, 2011) and Baroni and Zamparelli (2010) make use of indirect supervision: They use the actual contexts in which a phrase occurs in a corpus as the target representation to learn parameters for their models. Zanzotto et al. (2010) also use indirect supervision, but with dictionary definitions as a source. Evaluation of phrase models is mostly on the data by Mitchell and Lapata, with a few exceptions: Landauer and Dumais (1997) predict textual coherence, and Wu and Schuler (2011) integrate the construction of phrase vectors with statistical parsing, in order to generalize lexicalized parsing from headwords to vectors.

TOWARDS VECTOR-BASED SENTENCE SEMANTICS

We now turn to approaches with an ambitious aim: building a vector space model that can serve as a general-purpose semantics for natural language sentences (Clark and Pulman 2007; Baroni and Zamparelli 2010; Coecke et al. 2010; Preller and Prince 2010; Grefenstette and Sadrzadeh 2011). A central problem for these approaches is the encoding of sentence structure. Approaches that ignore word order may still perform well on the task of predicting phrase similarity, and for two-word phrases they do (Mitchell and Lapata 2010). But a general-purpose semantics must be able to distinguish ''dog bites man'' and ''man bites dog'', because human hearers can distinguish the two phrases and draw different conclusions from them, and also because some Natural language processing applications require a semantics that can distinguish them. The current prevalent framework uses logic for representing sentence semantics. We briefly sketch this framework before we discuss vector-based semantics.

LOGIC-BASED SEMANTICS

In logic-based semantics, the meaning of a sentence lies in the conditions under which it is true. A sentence is represented as a logic formula, which is interpreted with respect to a *model*, an abstract representation of a situation or setting (Montague 1970; Bos 2011). For example, the sentence ''a pink duck waddles'' can be represented in first-order logic as

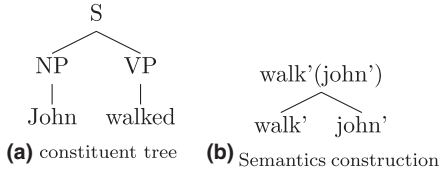
$$\exists x.duck'(x) \wedge pink'(x) \wedge waddle'(x)$$

This formula is true in a given model if the model contains an entity that is a duck, is pink, and belongs to the set of entities that waddle. Once a sentence or text is transformed to a logic-based representation, theorem provers can be used to draw inferences from it. The representation for a complex expression such as ''John walked'' is constructed based on the principle of compositionality (Frege's principle), which states that the meaning of a phrase is determined by the meanings of its components and the relations between them. As Figure 9 illustrates, the representation of ''John walked'' is constructed out of the representations for ''John'' and ''walked'' according to their syntactic relation. The sentence is represented as *walk' (john')* and not *john'(walk')*, matching the semantic types of the component phrases (Figure 9c).

Logic-based semantics has a long history in philosophy of language, artificial intelligence, and computational linguistics. It can easily represent even complex sentence structure. Proponents of a vector space semantics argue that logic-based semantics has no notion of either word similarity or phrase similarity, and typically has an impoverished lexical representation in which the representation of ''walk'' is just *walk'*. Both of these are problems that vector space approaches can solve. Approaches that work towards a vector-based semantics usually adopt the principle of compositionality for constructing phrase representations out of representations for their parts. They choose either syntactic or semantic types as the basis for the way that vector space representations for smaller expressions are combined to form larger phrases.

EXPLICIT SYNTACTIC STRUCTURE

Syntactic structure can be encoded explicitly in a vector space representation, such that the original syntactic structure can be read off the vector, or it can be implicit. In that case, it is used in the construction, but the original syntactic structure can no longer be identified from the resulting vector. The model of Smolensky (1990) encodes syntax explicitly, based on the insight that in a matrix (a second-order tensor) that is large enough, the whole syntactic structure can be embedded. Smolensky notes that a syntactic structure can be characterized by a set of pairs *role:filler*. He represents each such pair as a tensor product $\vec{role} \otimes \vec{filler}$, and the whole predicate as the sum of the role:filler vectors. (See Figure 6c for an illustration of the tensor product). It is possible to choose the



(a) constituent tree **(b)** Semantics construction

expression	type
john'	– entity
walk'	– function mapping entities to truth values
walk'(john')	– truth value

(c) Semantic types

Fig 9. Semantics construction: The semantic representation for the sentence is walk' (john'). It is constructed from the semantics of ''walked'' and ''John'' according to the syntactic structure and the semantics construction rules.

representations in such a way that the vectors for different role:filler pairs occupy different parts of the resulting matrix, for example by having them be linearly independent. In that case, the syntactic structure can be read off the matrix. Clark and Pulman (2007) propose a similar model. A downside of these approaches is that the dimensionality of the tensor for the sentence depends on the size of the syntax graph. Also, phrase similarity can only be computed for sentences that have the same structure.

IMPLICIT SYNTACTIC STRUCTURE

Baroni and Zamparelli (2010) focus on adjective/noun pairs and use semantic types to guide the construction of phrase vectors: The meaning of a noun is a set of entities (for example, the set of all ducks) and an adjective is a function from the meaning of a noun to the meaning of the modified noun (for example, the set of all pink ducks). Baroni and Zamparelli represent adjectives as matrices that map the nouns vectors to new vectors for the compound phrase. Syntactic structure is implicit here: It is used in the construction, but the resulting phrase vectors are indistinguishable from noun vectors, and in fact can be compared to them for similarity.

The approach of Coecke et al. (2010) is not specific to any part of speech. They also construct phrase representations compositionally, but based on syntactic rather than semantic types. Representations of words and phrases live in tensor spaces, the order of which depends on the syntactic type. This framework can be used either for sentence similarity predictions, or to encode truth-conditional semantics. For the latter case, a model (in the logic-based semantics sense of the word) is encoded in vector spaces. The representation of an intransitive verb like ''waddle'' is a vector in which each dimension stands for an entity, and a value of 1 means that that particular entity belongs to the set of waddlers in the model. The representation of a sentence is a vector in a one-dimensional space, in which a value of 1 indicates *true*, and 0 is *false*. The approach of Coecke et al. (2010) is vector-based, but was only recently implemented in distributional terms by Grefenstette and Sadrzadeh (2011). In their setting, nouns are vectors, and a transitive verb is a matrix that encodes seen pairs of a subject and an object, or rather their vectors. The meaning of a phrase ''subj verb obj'' is computed as

$$(\vec{subj} \otimes \vec{obj}) \odot \vec{verb}$$

for example $(\vec{john} \otimes \vec{duck}) \odot \vec{see}$ for the sentence ''John sees ducks''. Subject and object vectors are combined using tensor product, and then used to filter entries of the verb matrix through component-wise multiplication.

COMPARING ARBITRARY-LENGTH SENTENCES FOR SIMILARITY

Socher et al. (2011) represent the meaning of a sentence not as a single vector but as a tree of vectors, one for each constituent. To compute similarity between sentences, they consider all pairwise similarities of any constituent of the first sentence with any constituent of the second. This yields a matrix of similarities. A classifier is then used to judge whether such a matrix stands for two sentences that are paraphrases.

An important problem for vector-based semantics is how it will scale up: Can fixed-length vectors provide representations for sentences of arbitrary complexity in a way that still offers adequate, fine-grained similarity predictions? The tree-of-vectors approach provides a solution in which the size of sentence representations can differ, and it is still

possible to compare sentences of different length. Socher et al. (2011) evaluate their approach on corpus sentences of arbitrary length and complexity, rather than manually selected low-complexity phrase pairs as other approaches.

OPEN PROBLEMS

Many important problems are still open for vector-based semantics. One is the scaling problem discussed in the previous paragraph. Another problem concerns function words: What is the vector space representation going to be for determiners, or for prepositions? Preller and Sadrzadeh (2011) make an important first step in this direction by proposing an encoding of ''not''. A third open problem is predicting similarity for sentences that do not share the same structure, a problem mentioned above for Smolensky (1990) but shared by other approaches. For instance, the Grefenstette and Sadrzadeh (2011) model cannot compare two sentences in which one has a transitive and the other an intransitive verb, like ''the vase broke'' and ''John broke the vase'': One will be represented by a vector and the other by a matrix.

COMBINING LOGIC-BASED SEMANTICS AND VECTOR SPACES

Garrette et al. (2011) have as their aim not to replace logic-based representations with vector spaces, but to combine the two, arguing that the frameworks are orthogonal in their abilities. Garrette et al. use a vector space model to predict paraphrases based on word similarity, inject the results into a logic-based representation, and use probabilistic reasoning to draw inferences from the joint clause set. But this approach, too, only constitutes a first step: It does not make use of phrase similarity, and it suffers from scaling problems introduced by the probabilistic reasoning framework.

CONTEXT-THEORETIC SEMANTICS

Clarke (2007, 2012) proposes a theoretical framework that replaces concrete corpora with a generative corpus model that can assign probabilities to arbitrary word sequences (documents). This eliminates the sparseness problem of finite corpora, such that both words and larger phrases can be given distributional representations. The vector for any word or phrase x lists the probabilities of all possible sequences uxv. That is, the dimensions are pairs (u, v) of a left and a right context, and the space is infinite-dimensional. Clarke also combines vector spaces and logic-based semantics by proposing a space in which the dimensions correspond to logic formulas. A word or phrase x is linked to formulas for sequences uxv in which it occurs, and each formula F is generalized to other formulas G that entail F. But this approach, again, is only a first step; it is not worked out how the representation could be used for inferences.

SIMILARITY VERSUS IDENTITY

Should a general-purpose semantics be based on semantic similarity or on truth conditions? This fundamental question has been raised before, in a debate on symbolic versus connectionist representations (e.g., Horgan and Tienson 1989, 1990; Fodor and Lepore 1992, 1999; Fodor 1997; Churchland 1998; Doumas and Hummel 2005). Fodor and Lepore (1999) phrase this question as one of content identity versus similarity and argue that if the aim is to determine whether ''Nixon is dead'', it is not useful to test whether someone

similar to Nixon is in a bad state of health. There is also a third possibility. Maybe what is needed is a framework that has a notion of both identity and similarity, and a way of deciding when, and how much, to generalize to similar entities and propositions.

Short Biography

Katrin Erk's area of research is computational linguistics, focusing on lexical semantics. She has worked on the corpus annotation with semantic roles, and on the induction of lexical knowledge from corpus data, in particular on learning selectional preferences. Her recent focus has been on word meaning representations. She is developing novel computational models of word meaning that do not search for a single best sense and that instead describe varying degrees of similarity between uses of a word. Her 2009 co-authored paper ''Investigations on Word Senses and Word Usages'' discusses a pilot study on word meaning annotation that draws on methods from psychological experiments, using multiple annotators for annotation on a scale. In her 2010 paper, ''What is word meaning, really? (And how can distributional models help us describe it?)'', she lays out evidence from psychology and lexicography for word meaning representations with ''fuzzy boundaries'' and argues for the use of distributional models to represent word meaning without the use of dictionary senses. Before coming to the University of Texas at Austin, where she currently teaches, Katrin Erk held a postdoc position in the Computational Linguistics Institute at Saarland University, Germany. She holds a diploma in Computer Science from Koblenz University, Germany, and a Doctorate in Engineering from Saarland University, Germany.

Acknowledgement

This work is funded by NSF grant IIS 0845925 ''CAREER: Word meaning: beyond dictionary senses''. We thank Marco Baroni, Daoud Clarke, Louise McNally, Sebastian Padó, and the anonymous reviewers for helpful discussion.

Notes

* Corresponding address: Katrin Erk, Department of Linguistics, The University of Texas at Austin, Calhoun Hall 512 1 University Station B5100 Austin, TX 78712, USA. E-mail: katrin.erk@mail.utexas.edu

[1] The terms *distributional model* and *vector space model* are mostly used as synonyms, but they can be distinguished. We call a model ''distributional'' if it represents a word or phrase through its observed contexts, and a ''vector space model'' if it represents meaning in a high-dimensional space. While many models that we discuss are both distributional and based on vector space, we also look at some where only one of the two terms apply.

[2] The features *coord-1 raspberry* and *coord raspberry* indicate that the corpus contained ''strawberry and raspberry'' as well as ''raspberry and strawberry''.

[3] Like the model of Jones and Mewhort (2007) the model of Baroni and Lenci (2010) can address syntactic phenomena. However, in Jones and Mewhort's model, word order knowledge *emerges* from raw text training data, while Baroni and Lenci use a dependency parser.

[4] We have discussed both vector addition and component-wise multiplication in Section 4 as models for constructing a representation for a single word in a sentence context. But they are better viewed as phrase models. As word occurrence models, they assign the same vector – and thus the same meaning – to all the words in the phrase, which is counterintuitive.

Works Cited

Almuhareb, A., and M. Poesio. 2004. Attribute-based and value-based clustering: an evaluation. Proceedings of EMNLP, 158–65, Barcelona, Spain: Association for Computational Linguistics.

Andrews, M., G. Vigliocco, and D. Vinson. 2009. Integrating experiential and distributional data to learn semantic representations. Psychological Review 116(3). 463–98.
Baroni, M., and A. Lenci. 2010. Distributional memory: a general framework for corpus-based semantics. Computational Linguistics 36(4). 673–721.
——, B. Murphy, E. Barbu, and M. Poesio. 2010. Strudel: a corpus-based semantic model based on properties and types. Cognitive Science 34(2). 222–54.
——, and R. Zamparelli. 2010. Nouns and vectors, adjectives are matrices: representing adjective-noun constructions in semantic space. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 1183–93. Cambridge, MA: Association for Computational Linguistics, October.
Barsalou, L. W. 2008. Grounded cognition. Annual Review of Psychology 59(1). 617–45.
Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. Latent dirichlet allocation. Journal of Machine Learning Research 3. 993–1022.
Bos, J. A. 2011. Survey of computational semantics: representation inference and knowledge in wide-coverage text understanding. Language and Linguistics Compass 5(6). 336–66.
Bruni, E., G. B. Tran, and M. Baroni. 2011. Distributional semantics from text and images. Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics, 22–32. Edinburgh, UK: Association for Computational Linguistics, July.
Budanitsky, A., and G. Hirst. 2006. Evaluating Wordnet-based measures of semantic distance. Computational Linguistics 32. 13–47.
Burgess, C., and K. Lund. 1997. Modelling parsing constraints with high-dimensional context space. Language and Cognitive Processes 12. 177–210.
Churchland, P. M. 1998. Conceptual similarity across sensory and neural diversity: the Fodor/lepore challenge answered. Journal of Philosophy 95. 5–32.
Clarke, D. 2007. A context-theoretic framework for compositionality in distributional semantics. Computational Linguistics 38(1). 41–71.
Clark, S., and S. Pulman. 2007. Combining symbolic and distributional models of meaning. Proceedings of the AAAI Spring Symposium on Quantum Interaction, 52–5. Stanford, CA: Association for the Advancement of Artificial Intelligence.
Clarke, D. 2007. Context-theoretic semantics for natural language: an algebraic framework. Sussex: University of Sussex.
Coecke, B., M. Sadrzadeh, and S. Clark. 2010. Mathematical foundations for a compositional distributed model of meaning. Lambek Festschrift Linguistic Analysis, 36. 345–84.
Deerwester, S., S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. 1990. Indexing by Latent semantic analysis. Journal of the Society for Information Science 41(6). 391–407.
Deschacht, K. and M. -F. Moens. 2009. Semi-supervised semantic role labeling using the Latent words language model. Proceedings of EMNLP, 21–9, Singapore: Association for Computational Linguistics.
Devereux, B., N. Pilkington, T. Poibeau, and A. Korhonen. 2009. Towards unrestricted large-scale acquisition of feature-based conceptual representations from corpus data. Research on Language and Computation 7. 137–70.
Dinu, G., and M. Lapata. 2010. Measuring distributional similarity in context. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 1162–72. Cambridge, MA: Association for Computational Linguistics, October.
Doumas, L. A. A., and J. E. Hummel. 2005. Modeling human mental representations: what works what doesnt, and why. The Cambridge handbook of thinking and reasoning, ed. by K. J. Holyoak and R. G. Morrison, 73–91. Cambridge, UK: Cambridge University Press.
Erk, K. 2009. Representing words as regions in vector space. Proceedings of CoNLL, Boulder, CO: Association for Computational Linguistics.
——. 2010. What is word meaning, really? (and how can distributional models help us describe it?). Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics, 17–26. Uppsala, Sweden: Association for Computational Linguistics.
——, and S. Padó. 2008. A structured vector space model for word meaning in context. Proceedings of EMNLP, Honolulu, HI: Association for Computational Linguistics.
——, and ——. 2010. Exemplar-based models for word meaning in context. Proceedings of ACL, Uppsala, Sweden: Association for Computational Linguistics.
Feng, Y., and M. Lapata. 2010. Visual information in semantic representation. Human language technologies: the 2010 Annual Conference of the North American chapter of the Association for Computational Linguistics, 91–9. Los Angeles, California: Association for Computational Linguistics.
Firth, J. R. 1957. A synopsis of linguistic theory 1930–1955. Studies in linguistic analysis, 1–32. Oxford: England: Blackwell Publishers.
Fodor, J. 1997. Connectionism and the problem of systematicity: why smolensky's solution still doesn't work. Cognition 62. 109–19.
——, and E. Lepore. 1992. Holism: a shopper's guide. Cambridge: Blackwell.

——, and ——. 1999. All at sea in semantic space: Churchland on meaning similarity. Journal of Philosophy 96. 381–403.
Gärdenfors, P. 2004. Conceptual spaces. Cambridge, MA: MIT press.
Garrette, D., K. Erk, and R. Mooney. 2011. Integrating logical representations with probabilistic information using markov logic. Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011), 105–14. Oxford, UK: Special Interest Group on Computational Semantics of the Association for Computational Linguistics.
Gorman, J., and J. R. Curran. 2006. Scaling distributional similarity to large corpora. ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL, Sydney.
Grefenstette, E. and M. Sadrzadeh. 2011. Experimental support for a categorical compositional distributional model of meaning. Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 1394–404. Edinburgh, Scotland, UK: Association for Computational Linguistics, July.
Guevara, E. 2010. A regression model of adjective-noun compositionality in distributional semantics. Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics, 33–7. Uppsala, Sweden: Association for Computational Linguistics July.
——. 2011. Computing semantic compositionality in distributional semantics. Proceedings of IWCS-9, Oxford, UK: Special Interest Group on Computational Semantics of the Association for Computational Linguistics.
Hanks, P. 2000. Do word meanings exist? Computers and the Humanities 34. 205–15.
Harris, Z. 1954. Distributional structure. Word 10. 146–62.
Hartung, M., and A. Frank. 2011. Assessing interpretable, attribute-related meaning representations for adjective-noun phrases in a similarity prediction task. Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics, 52–61. Edinburgh, UK: Association for Computational Linguistics, July.
Hindle, D. 1990. Noun classification from predicate-argument structures. Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (ACL'90), Pittsburg, Pennsylvania: Association for Computational Linguistics.
Hofmann, T. 1999. Probabilistic latent semantic indexing. Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval, University of Berkeley, CA.
Horgan, T., and J. Tienson. 1989. Representations without rules. Philosophical Topics 17. 147–74.
——, and ——. 1990. Soft laws. Midwest studies in Philosophy 15. 256–79.
Jones, M. N., and D. J. K. Mewhort. 2007. Representing word meaning and order information in a composite holographic lexicon. Psychological Review 114. 1–37.
Kilgarriff, A. 1997. I don't believe in word senses. Computers and the Humanities 31. 91–113.
Kintsch, W. 2001. Predication. Cognitive Science 25. 173–202.
——. 2007. Meaning in context. Handbook of Latent semantic analysis, ed. by T. K. Landauer, D. McNamara, S. Dennis, and W. Kintsch, 89–105. Mahwah, NJ: Erlbaum.
Kremer, G., and M. Baroni. 2010. Predicting cognitively salient modifiers of the constitutive parts of concepts. Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics, 54–62. Uppsala, Sweden: Association for Computational Linguistics, July.
Landauer, T., and S. Dumais. 1997. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review 104(2). 211–40.
Leong, C. W., and R. Mihalcea. 2011. Measuring the semantic relatedness between words and images. Proceedings of IWCS-9, Oxford, UK: Special Interest Group on Computational Semantics of the Association for Computational Linguistics.
Lin, D. 1998. Automatic retrieval and clustering of similar words. COLING-ACL98, Montreal, Canada: Association for Computational Linguistics.
——, and P. Pantel. 2001. Discovery of inference rules for question answering. Natural Language Engineering 7(4). 343–60.
Louwerse, M. M. 2007. Symbolic or embodied representations: a case for symbol interdependency. Handbook of Latent semantic analysis, ed. by T. Landauer, D. McNamara S. Dennis. and W. Kintsch, 107–20. Mahwah, NJ: Erlbaum.
——. 2010. Symbol interdependency in symbolic and embodied cognition. Topics in Cognitive Science 3. 273–302.
Lowe, W. 2001. Towards a theory of semantic space. Proceedings of the Cognitive Science Society, 576–81. Mahwah, NJ: Erlbaum.
Lund, K., C. Burgess, and R. A. Atchley. 1995. Semantic and associative priming in high-dimensional semantic space. Proceedings of the Seventeenth Annual Meeting of the Cognitive Science Science Society, 660–5. Cognitive Science Society, University of Pittsburgh.
Manning, C. D., P. Raghavan, and H. Schütze. 2008. Introduction to information retrieval. Cambridge: Cambridge University Press.

McCarthy, D., R. Koeling, J. Weeds, and J. Carroll. 2004. Finding predominant senses in untagged text. Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL'04), 280–7. Barcelona, Spain: Association for Computational Linguistics.

——, and R. Navigli. 2009. The English Lexical substitution task. Language Resources and Evaluation 43(2). 139–59. Special Issue on Computational Semantic Analysis of Language: SemEval-2007 and Beyond

McDonald, S., and M. Ramscar. 2001. Testing the distributional hypothesis: the influence of context on judgements of semantic similarity. Proceedings of the Cognitive Science Society, 611–6. Cognitive Science Society, Edinburgh, Scotland.

McRae, K., G. S. Cree, M. S. Seidenberg, and C. McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. Behavior Research Methods 37(4). 547–59.

——, M. Spivey-Knowlton, and M. Tanenhaus. 1998. Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. Journal of Memory and Language 38. 283–312.

Mitchell, J., and M. Lapata. 2008. Vector-based models of semantic composition. Proceedings of ACL, Columbus, OH: Association for Computational Linguistics.

——, and ——. 2010. Composition in distributional models of semantics. Cognitive Science 34(8): 1388–429.

Montague, R. 1970. English as a formal language. Linguaggi nella societá e nella tecnica, ed. by B. Visentini, 189–224. Milan: Edizioni di Comunitá.

Moon, T., and K. Erk. forthcoming. An inference-based model of word meaning in context as a paraphrase distribution. ACM Transactions on Intelligent Systems and Technology special issue on paraphrasing to appear.

Padó, S., and M. Lapata. 2007. Dependency-based construction of semantic space models. Computational Linguistics 33(2). 161–99.

——, U. Padó, and K. Erk. 2007. Flexible, corpus-based modelling of human plausibility judgements. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), 400–9. Prague, Czech Republic: Association for Computational Linguistics, June.

Preller, A., and V. Prince. 2010. A quantum logic for compositional functional semantics and distributional semantic models. ESSLLI workshop on compositionality and distributional semantic models. Copenhagen: Association for Logic, Language and Information.

——, and M. Sadrzadeh. 2011. Bell states and negative sentences in the distributed model of meaning. Electronic Notes in Theoretical Computer Science 270(2). 141–53.

Ravichandran, D., P. Pantel, and E. Hovy. 2005. Randomized algorithms and NLP: using locality sensitive hash functions for high speed noun clustering. Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-05), 622–9. Ann Arbor, MI: Association for Computational Linguistics.

Reddy, S., I. P. Klapaftis, D. McCarthy, and S. Manandhar. 2011. Dynamic and static prototype vectors for semantic composition. Proceedings of The 5th International Joint Conference on Natural Language Processing 2011 (IJCNLP 2011), Chiang Mai, Thailand: Association for Computational Linguistics, November 2011.

Reisinger, J., and R. J. Mooney. 2010. Multi-prototype vector-space models of word meaning. Proceeding of NA-ACL. Los Angeles, CA: Association for Computational Linguistics.

Rubenstein, H., and J. Goodenough. 1965. Contextual correlates of synonymy. Computational Linguistics 8. 627–33.

Schütze, H. 1998. Automatic word sense discrimination. Computational Linguistics 24(1). 97–123.

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

Snow, R., D. Jurafsky, and Y. N. Andrew. 2006. Semantic taxonomy induction from heterogenous evidence. Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, 801–8. Sydney, Australia: Association for Computational Linguistics.

Socher, R., E. H. Huang, J. Pennin, A. Y. Ng, and C. D. Manning. 2011. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. Advances in Neural Information Processing Systems, ed. by J. Shawe-Taylor, R. S. Zemel, P. Bartlett, F. C. N. Pereira and K. Q. Weinberger, 24, 801–9. Granada: Neural Information Processing Systems Foundation.

Thater, S., G. Dinu, and M. Pinkal. 2009. Ranking paraphrases in context. Proceedings of the ACL Workshop on Applied Textual Inference, Singapore: Association for Computational Linguistics.

——, H. Fürstenau, and M. Pinkal. 2010. Contextualizing semantic representations using syntactically enriched vector models. Proceedings of ACL, Uppsala, Sweden: Association for Computational Linguistics.

——, ——, and ——. 2011. Word meaning in context: a simple and effective vector model. Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP), Chiang Mai, Thailand: Association for Computational Linguistics.

Turney, P. D. 2006. Similarity of semantic relations. Computational Linguistics 32(3). 379–416.

——, and P. Pantel. 2010. From frequency to meaning: vector space models of semantics. Journal of Artificial Intelligence Research 37. 141–88.

Van de Cruys T., T. Poibeau, and A. Korhonen. 2011. Latent vector weighting for word meaning in context. Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 1012–22. Scotland, UK, Edinburgh: Association for Computational Linguistics, July.

Vigliocco, G., D. P. Vinson, W. Lewis, and M. F. Garrett. 2004. Representing the meanings of object and action words: the featural and unitary semantic space hypothesis. Cognitive Psychology 48. 422–88.

Washtell, J. 2010. Expectation vectors: a semiotics inspired approach to geometric lexical-semantic representation. Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics, 45–50. Uppsala, Sweden: Association for Computational Linguistics, July.

——. 2011. Compositional expectation: a purely distributional model of compositional semantics. Proceedings of IWCS-9, Oxford, UK: Special Interest Group on Computational Semantics of the Association for Computational Linguistics.

Wittgenstein, L. 1997. Philosophical investigations. Oxford, England: Blackwell Publishers. Original work published 1953.

Wu, S., and W. Schuler. 2011. Structured composition of semantic vectors. Proceedings of IWCS-9, Oxford, UK: Special Interest Group on Computational Semantics of the Association for Computational Linguistics.

Zanzotto, F. M., I. Korkontzelos, F. Fallucchi, and S. Manandhar. 2010. Estimating linear models for compositional distributional semantics. Proceedings of the 23rd International Conference on Computational Linguistics (COLING). Beijing: Association for Computational Linguistics.