Annual Review of Linguistics

Distributional Models of Word Meaning

Alessandro Lenci

Computational Linguistics Laboratory, University of Pisa, Pisa 56126, Italy;
email: alessandro.lenci@unipi.it

Annu. Rev. Linguist. 2018. 4:151–71

First published as a Review in Advance on October 4, 2017

The *Annual Review of Linguistics* is online at linguist.annualreviews.org

https://doi.org/10.1146/annurev-linguistics-030514-125254

Keywords

distributional semantics, vector space models, linguistic contexts, lexicon, semantic similarity, compositionality

Abstract

Distributional semantics is a usage-based model of meaning, based on the assumption that the statistical distribution of linguistic items in context plays a key role in characterizing their semantic behavior. Distributional models build semantic representations by extracting co-occurrences from corpora and have become a mainstream research paradigm in computational linguistics. In this review, I present the state of the art in distributional semantics, focusing on its assets and limits as a model of meaning and as a method for semantic analysis.

1. DISTRIBUTIONAL SEMANTICS: FROM USAGE TO MEANING

Distributional semantics (DS), also known as vector space semantics, is a usage-based model of meaning, based on the assumption that the statistical distribution of linguistic items in context plays a key role in characterizing their semantic behavior. Its main focus is the lexicon: DS is primarily an empirical method for the analysis of lexical meaning (but see Section 5.2 for distributional models of compositional semantics). DS offers both a model to represent meaning and computational methods to learn such representations from language data. Given the ever-increasing availability of digital texts, distributional models can rely on huge amounts of empirical evidence to characterize the semantic properties of lexemes. Distributional representations are built from text corpora as samples of language usage and offer new ways to investigate the interplay between meaning and contexts, and to tackle the dynamicity and plasticity of meaning.

In this review, I present the state of the art in DS mainly from a linguistic perspective. Therefore, I focus on its assets (and limits) as a model of meaning and as a method for semantic analysis, leaving aside its applications in natural language processing (NLP). Jurafsky & Martin (2008) and Turney & Pantel (2010) offer surveys on DS from a more computational perspective; for other general introductions to DS, see Lenci (2008), Erk (2012), and Clark (2015).

2. THE DISTRIBUTIONAL HYPOTHESIS

The theoretical foundation of DS has become known as the distributional hypothesis (DH): Lexemes with similar linguistic contexts have similar meanings. The root of the DH lies in the distributionalism advocated by American structural linguists, in particular by Harris (1954, p. 156), who argued that "difference of meaning correlates with difference of distribution."

Distributional models of meaning have also been explored in psychology and cognitive science. A strenuous supporter of the importance of linguistic distributions in shaping semantic representations was Miller (1967), who considered Harris's distributional analysis a method to provide an empirical foundation for the notion of semantic similarity (see the sidebar titled Historical Notes). A definition of semantic similarity in distributional terms was more explicitly theorized by Miller & Charles (1991, p. 3), who conceived it as a "function of the contexts in which words are used." DS is not only a method for lexical analysis but also a theoretical framework to build computational models of semantic memory (McRae & Jones 2013) that assume "a formal cognitive mechanism to learn semantics from repeated episodic experience in the linguistic environment (typically a text corpus)" (Jones et al. 2015, p. 239).

An essential contribution to the development of distributional semantics has come from the vector space model in information retrieval (Salton et al. 1975), which represents a collection of documents with a matrix whose rows are vectors corresponding to lexical items and whose columns

HISTORICAL NOTES

One of the first appearances of the term "distributional semantics" is by Garvin (1962), who used it to refer to a research program in machine translation inspired by Harris's distributionalism. The development of DS was also indirectly but strongly influenced by the later writings of Wittgenstein (1953) and by the contextual view of meaning advocated by Firth (1957), which prompted research on collocations in corpus linguistics. Vector-based representations of meaning, like those later adopted in DS, were pioneered in psychology by Osgood (1952), who defined the semantic system as a semantic space of *n*-dimensional feature vectors representing concepts (however, the dimensions of Osgood's semantic spaces were not corpus based).

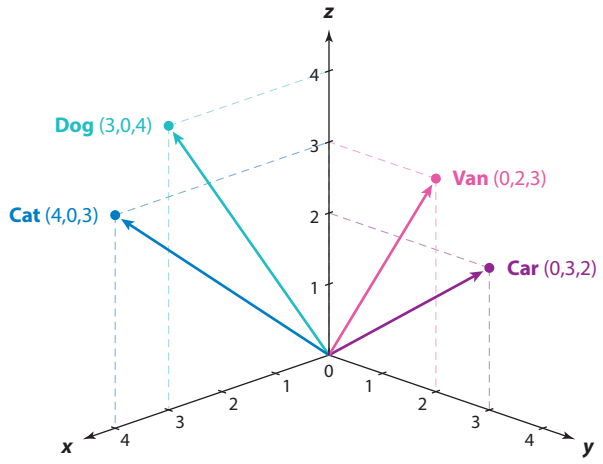

Figure 1

Distributional vectors of the lexemes *car*, *cat*, *dog*, and *van*.

are vectors corresponding to documents, and each matrix entry records the occurrences of a lexical item in a document. Since its conception, the vector space model has also been used to identify semantically associated words by measuring the similarity of their corresponding vectors. While DS continued to be pursued in information retrieval, it was virtually ignored in computational linguistics until the early 1990s, because of the dominance of formal and logic methods. The new empiricist turn and the emergence of statistical NLP, together with the availability of larger corpora and faster computers, favored a growing interest in DS, which has become a mainstream research paradigm in computational linguistics.

3. DISTRIBUTIONAL REPRESENTATIONS

The DH states that the semantic similarity of lexical items is a function of their distribution in linguistic contexts. Distributional representations operationalize this assumption by providing a mathematical encoding of the distributional properties of lexemes. The distributional representation of a lexical item is typically a distributional vector representing its co-occurrences with linguistic contexts—hence the name vector space semantics.

Vectors have geometrical interpretations: Vectors with n components define points (or arrows) in n-dimensional spaces. Therefore, distributional representations are geometrical representations of the lexicon in the form of a distributional vector space. The positions of lexemes in a distributional semantic space depend on their co-occurrences with linguistic contexts. **Figure 1** represents the lexemes *car*, *cat*, *dog*, and *van* in a three-dimensional vector space (vectors are marked in bold). Semantic representations are typically couched in symbolic terms and meanings are represented with symbols of some formal metalanguage (e.g., first-order logic, semantic networks, frames, feature structures). Symbolic semantic representations are therefore discrete and categorical. Distributional representations are instead graded and distributed, because information is encoded in the continuous values of vector dimensions.

Vector: a vector $\mathbf{v}$ is an ordered list of real numbers $(v_1, \ldots, v_n)$; v_i is the ith component of the vector

Distributed versus distributional: in distributed representations, information is distributed across vector dimensions; distributional vectors are distributed representations recording co-occurrences of lexemes with linguistic contexts

3.1. Context Types

Distributional representations differ with respect to the way linguistic contexts are defined (**Table 1**). The arguably most common type of context is the set of collocates of a target

Table 1 Examples of co-occurrences extracted from the same text fragment for the target *know* with different context types

Firth (1957): [*You shall* know *a word*] by the company it keeps![a]	
Context types	**Co-occurrences**
Undirected window-based collocate	*word*
Directed window-based collocate	⟨R, *word*⟩[b]
Dependency-filtered syntactic collocate	*word*
Dependency-typed syntactic collocate	⟨obj, *word*⟩[c]
Text region	Firth (1957)

[a] *You*, *shall*, and *a* are treated as stop words and not listed as collocates.
[b] R indicates that the context lexeme appears to the right of the target.
[c] The context lexeme is typed with the syntactic dependency (e.g., direct object) linking it to the target.

lexeme, that is, the "company" of context lexemes co-occurring with the targets (Firth 1957). The context lexemes are a subset of the corpus vocabulary, possibly coinciding with the targets. It is also customary to filter out words that are not informative from the semantic point of view. These so-called stop words include the most frequent lexemes in any corpus, such as grammatical function words.

The kind of co-occurrence relation between target and context lexemes determines different types of collocates and distributional representations. The window-based collocates of a target t are context lexemes that occur within a certain linear distance from t specified by a context window. This is the most common way to characterize lexical collocates and is directly derived from the Firthian tradition of collocation analysis. Undirected collocates do not distinguish between context lexemes appearing to the left and to the right of the target, whereas directed collocates treat lexemes appearing to the left and to the right of the target as different contexts. The size of the context window significantly affects the type of collocates that are extracted for a given target. No theoretical principle exists to determine the best context window, which is therefore a parameter to be fixed empirically. Window-based collocates are the most popular kind of linguistic contexts in DS, because they are extremely simple and fast to extract and do not require deep linguistic processing of the training corpus (even a simply tokenized text may suffice). By contrast, window-based collocates do not take into account linguistic structure, because context windows are viewed as sums of independent words, ignoring syntactic information. This characterization of linguistic contexts, known in the literature as the bag of words model, provides a shallow representation of their content. In Harris's (1954, p. 156) words, "language is not merely a bag of words."

The syntactic collocates of a target t are lexemes that have a syntactic relation with t. The standard way to identify syntactic collocates in DS is in terms of dependency relations. In dependency-filtered collocates, syntactic dependencies are used only to define the syntactic collocates, without entering into the specification of the contexts themselves: Identical lexemes linked to the target by different dependency relations are mapped onto the same context. In dependency-typed collocates, syntactic dependencies are instead encoded in the contexts. On one hand, syntactic collocates are attractive because they take into account the linguistic structure of contexts; on the other hand, they need to be extracted from dependency-parsed corpora. In general, the question of whether syntactic information provides a real advantage over bag-of-words representations of contexts is still open and highly dependent on the semantic task (Kiela & Clark 2014, Lapesa & Evert 2017).

The distributional properties of lexical items are also represented with the texts in which they occur. A (text) region is any uniquely identifiable text sample: book chapters, web pages, or simply

text portions of any fixed size (**Table 1**). This approach stems directly from the vector space model in information retrieval (see Section 2). Texts can be regarded as episodes (Landauer & Dumais 1997) that become associated with the words encountered therein. Lexemes are thus similar to the extent that they appear in the same episodes.

Matrix: a matrix $A_{m \times n}$ is an array of numbers with m rows and n columns; a_{ij} is the entry in the ith row and jth column of A

Zipf's law: the frequency of a word, $F(w)$, is inversely proportional to its rank, $r(w)$, given the constants C and a:

$$F(w) = \frac{C}{r(w)^a}$$

3.2. Building Distributional Representations

The basic method of building distributional vectors consists of the following procedure:

- co-occurrences between lexical items and linguistic contexts are extracted from a corpus and counted;
- the distribution of lexical items is represented with a co-occurrence matrix, whose rows correspond to target lexical items, columns to contexts, and the entries to their co-occurrence frequency;
- raw frequencies are then usually transformed into significance weights to reflect the importance of the contexts; and
- the semantic similarity between lexemes is measured with the similarity between their row vectors in the co-occurrence matrix.

Suppose we have extracted and counted the co-occurrences of the targets $T = \{bike, car, dog, lion\}$ with the context lexemes $C = \{bite, buy, drive, eat, get, live, park, ride, tell\}$ in a corpus. Their distribution is represented with the following co-occurrence matrix $M_{T \times C}$, in which $m_{t,c}$ is the co-occurrence frequency of t with c:

$$(1) \quad \begin{array}{c} \\ bike \\ car \\ dog \\ lion \end{array} \begin{array}{c} \begin{array}{ccccccccc} bite & buy & drive & eat & get & live & park & ride & tell \end{array} \\ \begin{pmatrix} 0 & 9 & 0 & 0 & 12 & 0 & 8 & 6 & 0 \\ 0 & 13 & 8 & 0 & 15 & 0 & 5 & 0 & 0 \\ 0 & 0 & 0 & 9 & 10 & 7 & 0 & 0 & 1 \\ 6 & 0 & 0 & 1 & 8 & 3 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

Matrix 1, with rows labeled with target lexemes and columns with context lexemes, is called a word-by-word co-occurrence matrix. A co-occurrence matrix whose columns are labeled with text regions is referred to as a word-by-region matrix.

One of the main tenets of DS is that co-occurrence frequency is a crucial clue to estimate the importance of distributional data in characterizing a target lexeme. However, weighting distributional pairs with their raw frequency is not the optimal solution. The problem lies in the fact that frequency distributions of lexemes follow Zipf's law and are highly skewed, with few very frequent lexical items and a large number of extremely rare ones. Distributional representations use various forms of weighting functions to overcome the problems of raw frequencies and to assign higher weights to co-occurrences that are more informative about the content of the target lexemes. The most common weighting function in DS is positive pointwise mutual information (PPMI) (Bullinaria & Levy 2007):

$$(2) \quad \text{PPMI}(t, c) = \max\left(0, \log_2 \frac{p(t, c)}{p(t)p(c)}\right).$$

PPMI measures how much the probability of a target–context pair estimated in the training corpus is higher than the probability we should expect if the target and the context occurred independently of one another. Matrix 3 contains the PPMI weights computed from the raw co-occurrence frequencies in matrix 1:

(3)

	bite	*buy*	*drive*	*eat*	*get*	*live*	*park*	*ride*	*tell*
bike	0	0.50	0	0	0	0	1.09	1.79	0
car	0	0.80	1.56	0	0	0	0.18	0	0
dog	0	0	0	2.01	0	1.65	0	0	2.16
lion	2.75	0	0	0	0.26	1.01	0	0	0

For other types of weighting functions used in DS, see Curran (2003), Evert (2008), Turney & Pantel (2010), and Kiela & Clark (2014).

The distributional similarity between two lexemes u and v is measured with the similarity between their distributional vectors $\mathbf{u}$ and $\mathbf{v}$. Once we have computed the pairwise distributional similarity between the targets, we can identify the k nearest neighbors of each target t, that is, the k lexical items with the highest similarity score with t. The cosine is the most popular measure of vector similarity in DS:

$$(4)\quad \cos(\mathbf{u},\mathbf{v}) = \frac{\mathbf{u}\cdot\mathbf{v}}{||\mathbf{u}||||\mathbf{v}||} = \frac{\sum_{i=1}^{n} u_i v_i}{\sqrt{\sum_{i=1}^{n} u_i^2}\sqrt{\sum_{i=1}^{n} v_i^2}}.$$

The cosine ranges from 1 for identical vectors to −1 (0, if the vectors do not contain negative values): The lower the cosine is, the lower the vector similarity will be. The following similarity matrix reports the cosines between the row vectors in matrix 3:

(5)

bike	1			
car	0.16	1		
dog	0	0	1	
lion	0	0	0.17	1
	bike	*car*	*dog*	*lion*

For other types of vector (dis)similarity measures, see Manning & Schütze (1999), Curran (2003), Bullinaria & Levy (2007), and Kiela & Clark (2014). Weeds et al. (2004) evaluated the effects of various measures on a word's distributional neighbors.

3.3. Explicit and Implicit Distributional Vectors

The distributional representations produced by co-occurrence matrices have three main properties: (*a*) Each vector dimension represents a (weighted) count of the co-occurrence of lexemes with a specific context, (*b*) vectors are high dimensional, as the number of contexts in language data tend to be very large, and (*c*) because of the Zipfian distribution of co-occurrence data, vectors are sparse, which means that most of their dimensions are zero. Following Levy & Goldberg (2014b), I refer to this kind of representation as an explicit distributional vector.

High-dimensional explicit vectors miss important generalizations in distributional data. Because they regard each context as a distinct feature, they do not take into account the fact that contexts may be very similar and strongly correlated with one another. Moreover, explicit vectors suffer from the fact that many possible co-occurrences remain unobserved in corpora regardless of their size, simply because of the skewed data distribution. In order to overcome these problems, lexemes can be represented with a different kind of distributional vector such that (*a*) the dimensions correspond to k latent features extracted from co-occurrences, (*b*) the number of latent features, typically on the order of a few hundreds, is much smaller than the original number of linguistic contexts, and (*c*) the vectors are dense, because most of the components are nonzero. Because there is no direct correspondence between vector features and linguistic contexts, I refer to such a representation as an implicit distributional vector.

SINGULAR VALUE DECOMPOSITION

The most popular feature extraction algorithm in distributional semantics is singular value decomposition (SVD), which factorizes an $m \times n$ co-occurrence matrix M into the product of three other matrices, where $z = \min(m, n)$:

$$M_{m\times n} = U_{m\times z}\Sigma_{z\times z}(V_{n\times z})^T.$$

The row vectors of U are called left singular vectors; the row vectors of V, right singular vectors. $\Sigma_{z\times z}$ is a square diagonal matrix containing singular values, sorted in descending order. The columns of the matrices U and V represent latent dimensions in the original data, ordered by the amount of variance they account for. By deleting all but the first k singular values and singular vectors (with k typically ranging between 100 and 300), we arrive at a new matrix $\hat{M}$ that is the best approximation of M in a reduced k-dimensional space, while retaining the majority of the variation in the data. This reduction is known as truncated SVD. To compute similarities between lexemes in the reduced space, we discard the V matrix and retain only the U and Σ matrices. Their product yields a reduced matrix $M'_{T\times D}$ with size $m \times k$:

$$M'_{T\times D} = U_{m\times k}\Sigma_{k\times k}.$$

The row vectors of M' are implicit distributional vectors with latent semantic dimensions $D = \{d_1, \ldots, d_k\}$. Alternatively, Σ can be dropped and the row vectors of U directly used to represent the targets. Levy et al. (2015a) show that the latter solution improves the quality of semantic representations.

Implicit vectors are created with dimensionality-reduction techniques that map the data in the high-dimensional space of linguistic contexts to a space of fewer latent dimensions. This process is also called feature extraction, because the dimensions of the reduced space are new features extracted from the original data. The main assumption is that co-occurrences collected from corpora are noisy data that hide more abstract semantic structures. Feature extraction aims to uncover such a latent structure and to eliminate the surface noise (Deerwester et al. 1990). Thus, instead of representing target lexemes using the linguistic contexts they co-occur with, we represent them in a latent semantic space of implicit vectors with a much smaller set of abstract features discovered in distributional data. The most common way to create implicit distributional representations is to map the co-occurrence matrix onto a reduced latent semantic space with a matrix reduction algorithm, such as singular value decomposition (SVD) (see the sidebar titled Singular Value Decomposition), principal components analysis (PCA), and nonnegative matrix factorization (NMF). For other ways to build implicit distributional vectors, see Section 4.

An important difference between explicit and implicit representations is the interpretability of their components. In explicit vectors, dimensions have a straightforward interpretation because there is a one-to-one correspondence between features and linguistic contexts. By contrast, it is usually difficult (if not impossible) to assign to each latent feature a clear semantic value.

4. DISTRIBUTIONAL SEMANTIC MODELS

The parameters to be determined when building distributional representations include the selection of target lexemes, the definition of context type, the choice of weighting scheme, the application of dimensionality reduction, and the choice of a vector similarity metric. A distributional semantic model (DSM) is a particular configuration of the parameters used to build distributional representations. The two major dimensions of separation between the various existing (and possible) models are (*a*) the type of context and (*b*) the method of learning distributional vectors (**Figure 2**).

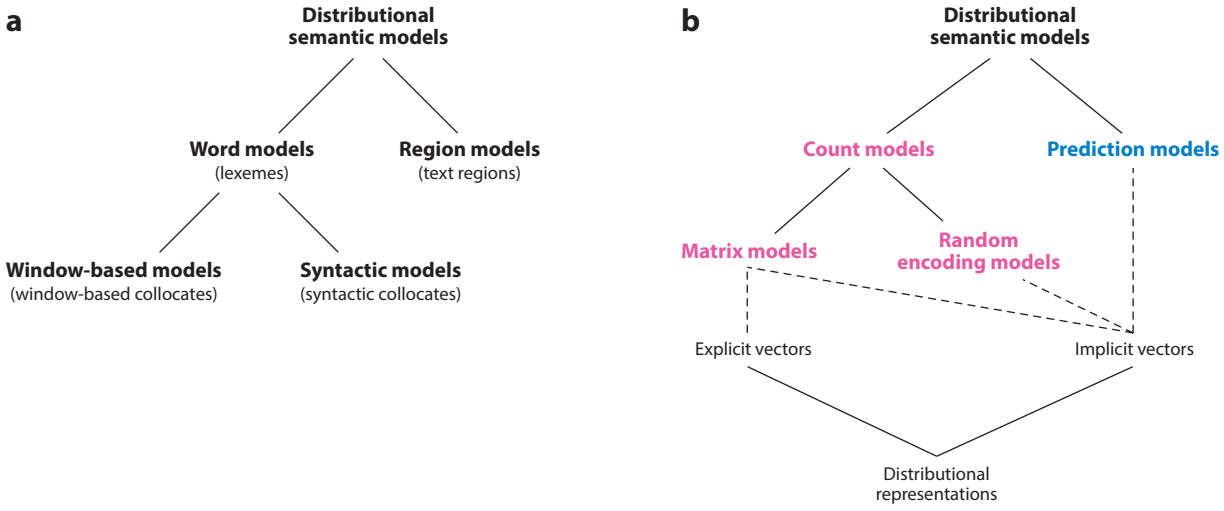

Figure 2

A classification of distributional semantic models based on (*a*) context types and (*b*) methods to build distributional vectors.

The choice of context is one of the most important parameters of DSMs, and it strongly affects the similarity relations they identify. A major contrast is between word models and region models, as they represent two radically different approaches to determining semantic similarity. Word models assume that two lexemes are similar if they tend to co-occur with the same collocates. By contrast, region models assume that two lexemes are similar if they tend to co-occur in the same texts. Region models therefore tend to identify semantic neighbors that are topically similar (i.e., belong to the same semantic domain or topic, like *car* and *race*) because they appear in documents about the same arguments. Word models tend to emphasize neighbors that are attributionally similar (i.e., share a number of common attributes, like *car* and *van*). The type of collocates can also affect the shape of the semantic space. For instance, experiments have shown that narrow context windows and syntactic collocates are best suited to capturing lexemes that are related by paradigmatic semantic relations (e.g., synonyms and antonyms) or that belong to the same taxonomic category (e.g., *violin* and *guitar*), because they share very close collocates (Sahlgren 2006, Bullinaria & Levy 2007, Van de Cruys 2008, Baroni & Lenci 2011, Bullinaria & Levy 2012, Kiela & Clark 2014, Levy & Goldberg 2014a). Conversely, collocates extracted with larger context windows are biased toward more associative semantic relations (e.g., *violin* and *music*), like region models.

The second dimension of variation among DSMs is the method to learn distributional representations. Matrix models (**Table 2**) are a rich family of DSMs that generalize the vector space model in information retrieval (see Section 2). They are a subtype of so-called count models (Baroni et al. 2014b), which learn the representation of a target lexeme by recording and counting its co-occurrences in linguistic contexts. Matrix models arrange distributional data into co-occurrence matrices. The matrix is a formal representation of the global distributional statistics extracted from the corpus. The weighting functions use such global statistics to estimate the importance of co-occurrences to characterize target lexemes. Matrix DSMs can represent lexemes with explicit or implicit distributional vectors. The latter are learned by mapping the co-occurrence matrix onto a new reduced space, typically using matrix factorization techniques such as SVD. Comparative evaluations of the effects of various parameter settings in matrix DSMs have been reported by Bullinaria & Levy (2007, 2012), Kiela & Clark (2014), and Lapesa & Evert (2014).

Table 2 The most common matrix distributional semantic models

Model	Description	Reference
Latent Semantic Analysis (LSA)	Word-by-region matrix, weighted with entropy and reduced with SVD	Landauer & Dumais (1997)
Hyperspace Analogue of Language (HAL)	Window-based model with directed collocates	Burgess (1998)
Dependency vectors (DV)	Syntactic model with dependency-filtered collocates	Padó & Lapata (2007)
Latent relational analysis (LRA)	Pair-by-pattern matrix reduced with SVD to measure relational similarity	Turney (2006)
Distributional memory (DM)	Target–link–context tuples formalized with a high-order tensor	Baroni & Lenci (2010)
Topic models	Word-by-region matrix reduced with Bayesian inference	Griffiths et al. (2007)
High-dimensional explorer (HiDEx)	Generalization of HAL with a larger range of parameter settings	Shaoul & Westbury (2010)
Global vectors (GloVe)	Word-by-word matrix reduced with weighted least-squares regression	Pennington et al. (2014)

Abbreviation: SVD, singular value decomposition.

Matrix models such as latent semantic analysis (LSA), hyperspace analogue of language (HAL), and dependency vectors (DV) directly implement the basic procedure to build distributional representations described in Section 3.2. Other DSMs, by contrast, introduce extensions and variants to the classical method. In order to measure relational similarity (i.e., word pairs linked by similar semantic relations, like *tulip–flower* and *dog–animal* sharing the hypernymy relation), latent relational analysis (LRA) uses a pair-by-pattern co-occurrence matrix, with rows corresponding to pairs of lexical items and columns to lexico-syntactic patterns linking them. Distributional memory (DM) is a generalized framework for DS that represents corpus data as ternary tuples (e.g., *boy–drink–milk*) formalized with a high-order tensor, from which different types of co-occurrence matrices are derived to address a wide range of semantic tasks. Topic models and global vectors (GloVe) introduce new methods to reduce the dimensionality of the co-occurrence matrix: Topic models use latent Dirichlet allocation (Blei et al. 2003), a Bayesian probabilistic model, and GloVe use weighted least-squares regression. Despite their popularity, matrix models have some drawbacks. Dimensional reduction can be computationally quite onerous when applied to very large matrices. Moreover, matrix DSMs lack incrementality because they rely on the global statistics collected in the co-occurrence matrix. If new distributional data are added, the whole semantic space must be rebuilt from scratch.

Tensor: a multiway array whose order is the number of indices needed to identify its elements; tensors generalize vectors (first-order tensors) and matrices (second-order tensors)

Random encoding models are another type of count DSMs: Rather than collect global co-occurrence statistics into a matrix and then optionally reduce them to a dense vector, such models directly learn low-dimensional implicit vectors by assigning each lexical item a random vector that is incrementally updated depending on the co-occurring contexts. The most famous DSM of this kind is random indexing (RI) (Kanerva et al. 2000, Sahlgren 2006), which accumulates distributional vectors in an online fashion. If we use lexemes as contexts, RI assigns random index vectors to the lexemes in the data, and adds the random index vectors of the neighboring lexemes to the distributional vector for the target:

$$(6) \quad \mathbf{t}_i \leftarrow \mathbf{t}_{i-1} + \sum_{j=-n, j\neq 0}^{n} \mathbf{c}_j,$$

where $\mathbf{t}_i$ is the target distributional vector at step i, n is the extension of the context window, and $\mathbf{c}_j$ is a sparse k-dimensional random index vector (with δ randomly placed +1s and −1s) that acts as a fingerprint of context term c_j. The bound encoding of the aggregate language environment (BEAGLE) model created by Jones & Mewhort (2007) is also based on random vectors, but it encodes sequence information with circular convolution and auxiliary random vectors that represent the position of the target word. Recchia et al. (2015) propose a simpler method to encode linear order that extends RI with random permutations of the random index vectors to reflect the position of context items.

Embedding: in the neural network literature, any information "embedded" in a low-dimensional vector space. Broadly speaking, all implicit distributional vectors are embeddings; in a narrow sense, word embeddings are distributional vectors built with neural networks

Prediction models are a new family of DSMs that take a radically different approach to learning distributional vectors. Instead of counting co-occurrences, prediction DSMs are neural network algorithms that directly create low-dimensional implicit distributional representations by learning to optimally predict the contexts of a target word. These representations are also commonly referred to as (neural) word embeddings. There are many variations of DSMs that use neural networks as processing models, ranging from simple recurrent networks (Elman 1990) to more complex deep architectures (Collobert & Weston 2008). The most popular neural DSM is the one implemented in the `word2vec` library, which uses the softmax function for predicting b given a (Mikolov et al. 2013a,b):

$$(7) \quad p(b|a) = \frac{\exp(\mathbf{b} \cdot \mathbf{a})}{\sum_{b' \in C} \exp(\mathbf{b}' \cdot \mathbf{a})},$$

where C is the set of context words and $\mathbf{b}$ and $\mathbf{a}$ are the vector representations for the context and target words, respectively. This general model has two versions: continuous bag of words (CBOW), which predicts a target word based on the context, and skip-gram with negative sampling (SGNS), which predicts the context on the basis of the current word. Various types of "linguistic regularities" have been claimed to be identifiable by neural embeddings (Mikolov et al. 2013c). For instance, the fact that *king* and *queen* have the same gender relation as *man* and *woman* is represented in their embeddings' offsets, so that the vector of one word (e.g., **queen**) can be recovered from the representations of the other words by simple vector arithmetics (i.e., **king** − **man** + **woman**).

Despite the increasing popularity of neural embeddings, the question of whether they are really a breakthrough with respect to more traditional DSMs is far from resolved. The same linguistic regularities captured by embeddings are also captured by matrix models with explicit distributional vectors (Levy & Goldberg 2014b). Baroni et al. (2014b) report that prediction models outperform count models in various semantic tests. However, when the parameters of the latter are carefully tuned, no significant difference is observed between the two kinds of models (Levy et al. 2015a). Mandera et al. (2017) also show that count and prediction models produce very similar results. Moreover, when trained on smaller data sets some matrix models are even superior to neural embeddings, which become competitive only when trained on many more data (Sahlgren & Lenci 2016). Future research might reveal a clear advantage of neural models, but at present the two approaches do not substantially differ for the semantic aspects they are able to address. They are simply alternative ways to build distributional representations.

5. RESEARCH QUESTIONS IN DISTRIBUTIONAL SEMANTICS

DS is based on a simple assumption: Semantic representations of lexical items can be built by recording their distribution in linguistic contexts. However, whether statistical co-occurrences alone are enough to address deep semantic questions, or whether they merely provide a shallow proxy of lexical meaning, remains an open question. In other words, what is the real descriptive

and explanatory adequacy of distributional representations of meaning? I explore this issue by presenting some research themes that shed light on the potentialities and the current limits of distributional models of meaning.

5.1. Semantic Similarity and Relatedness

Modeling semantic similarity is one of the main success stories of distributional models. This is hardly surprising, because the DH itself is a claim about semantic similarity, which is inherently a graded notion and therefore an ideal benchmark for distributional representations. The primary outcome of DSMs is a continuous semantic space defined by mutual proximity relations among lexical items.

Semantic similarity is the most common and basic means of testing the performance of distributional models. DSMs are typically evaluated for (*a*) accuracy in multiple-choice synonym detection tasks like the one used in the Test of English as a Foreign Language (TOEFL) (Landauer & Dumais 1997) and (*b*) correlation with human similarity ratings collected in data sets such as the small RG65, used in the seminal study by Rubenstein & Goodenough (1965), and the much larger and more recent WordSim-353 (Finkelstein et al. 2001), MEN (Bruni et al. 2014), and SimLex-999 (Hill et al. 2015). Performance greatly varies depending on the data set and the model, but DSMs can achieve perfect accuracy on the TOEFL (Bullinaria & Levy 2012) and a Spearman correlation of 0.8 or better with similarity ratings. The performance of DSMs is often better than that obtained with measures based on manually designed lexical resources like WordNet (Agirre et al. 2009, Lofi 2015). DSMs also obtain very good results in semantic tasks that indirectly involve semantic similarity, such as categorizing nouns (Baroni & Lenci 2010, Riordan & Jones 2011), modeling semantic priming (Jones et al. 2006, Mandera et al. 2017), and predicting patterns of functional magnetic resonance imaging (fMRI) activation (Mitchell et al. 2008, Anderson et al. 2017). DS is routinely used in cognitive science to estimate semantic similarity, as an alternative to the direct elicitation of humans' ratings or to lexical resources (Keuleers & Balota 2015, Mandera et al. 2015).

These positive results notwithstanding, the relationship between semantic similarity and DSMs is much more complex and problematic than it appears prima facie. First, semantic similarity is itself a very vague notion. We must distinguish semantic similarity, sensu stricto, as a relation between words sharing similar semantic features, such as as *car* and *van*, from the broader notion of semantic relatedness between words that are strongly associated, like *car* and *driver* (Budanitsky & Hirst 2006). These two types of relations have very different properties. Yet they are barely distinguished by DSMs. Except for SimLex-999, the other data sets typically used for DSM evaluation contain ratings about semantic relatedness, rather than genuine semantic similarity. Hill et al. (2015) show that the performance of all DSMs significantly decreases when evaluated on SimLex-999, meaning that DSMs tend to identify broadly related neighbors, rather than similar ones.

Most of DSM evaluation has focused on nouns. SimVerb-3500 is a large data set with similarity ratings between verbs: Both count and prediction models achieve a very low correlation with human ratings (not exceeding 0.4), and are often outperformed by nondistributional similarity measures (Gerz et al. 2016). This finding reveals that DSMs show very uneven behavior depending on the area of the lexicon, and that current distributional representations might be inadequate to deal with the complexity of verb semantics.

An additional problem is that both semantic similarity and semantic relatedness are cover terms for various types of lexical relations: hypernymy, antonymy, meronymy, locative relations, and topical and other nonclassical relations (Morris & Hirst 2004). Moreover, these semantic relations have very different inferential properties (Murphy 2003). For instance, *John has a dog*

Table 3 Nearest neighbors in a CBOW ordered from left to right by similarity

Target	Neighbors[a]
car	*truck*, *vehicle*, *driving*, *garage*, *drive*, *jeep*, *windshield*, *driver*, *drove*, *bike*
smart	*dumb*, *clever*, *stupid*, *intelligent*, *pretty*, *enough*, *tough*, *you*, *think*, *cute*
eat	*hungry*, *eating*, *ate*, *eaten*, *eats*, *food*, *meal*, *starving*, *lunch*, *delicious*

[a]See **http://meshugga.ugent.be/snaut-english/** (Mandera et al. 2017).
Abbreviation: CBOW, continuous bag of words.

entails that *John has an animal*, because *animal* is a hypernym of *dog*, but does not entail that *John has a cat*, because *cat* is a cohyponym of *dog*. In general, the semantic neighbors identified by DSMs have multifarious relations with the target (**Table 3**), suggesting that DSMs provide a quite coarse-grained representation of lexical meaning. Semantic relation discrimination is an important area of research in DS, and some data sets, like BLESS (Baroni & Lenci 2011) and EVALution (Santus et al. 2015), were designed specifically to test DSMs on this task.

Most research in this area has focused on hypernym identification (Shwartz et al. 2017). Unsupervised approaches compute a score that is expected to be higher for hypernym pairs than for negative instances. Each measure exploits some variation of the DH on the basis of either feature inclusion (Kotlerman et al. 2010, Lenci & Benotto 2012) or feature informativeness (Rimell 2014, Santus et al. 2014a), aiming to capture the fact that hypernyms are semantically broader than hyponyms. Supervised models, by contrast, represent word pairs with a combination (e.g., concatenation or difference) of their distributional vectors, and train a classifier on the combined vectors to identify hypernyms (Baroni et al. 2012, Roller et al. 2014, Weeds et al. 2014). Supervised models typically achieve better results, but Levy et al. (2015b) cast doubts on the explanatory adequacy of such approaches.

Antonymy also represents a significant challenge for DSMs. Mohammad et al. (2013) have shown that synonyms and antonyms are indistinguishable in terms of their degree of distributional similarity (see the neighbors of *smart* in **Table 3**), because both tend to occur in similar contexts (Miller & Charles 1991). Current approaches to determining antonymy range from identifying contexts that are expected to be more discriminative of contrast (Turney 2008, Santus et al. 2014b) to using hybrid models in which DSMs are enriched with information extracted from lexical resources (Yih et al. 2012). The results are promising, but still not fully satisfactory.

In summary, although the DH is couched in terms of similarity, DSMs are actually more biased toward the much vaguer notion of semantic relatedness. The outcome of DSMs resembles a network of word associations, rather than a semantically structured space—an important weakness. Although these DSMs have proven useful in capturing various aspects of the mental lexicon, their limitations in properly distinguishing different semantic relations greatly impair the utility of DS in modeling logical inferences (Erk 2016). Whether more fine-grained semantic relations can be identified with purely distributional methods is an open research issue.

5.2. Beyond the Lexicon: Compositional Distributional Semantics

A central aspect of human semantic competence is the ability to compose lexical meanings to form the interpretation of a potentially unlimited number of complex linguistic expressions. Formal semantics provides a rigorous logico-mathematical model for semantic compositionality that computes the truth conditions of a sentence as a function of (*a*) the interpretation of its lexical items and (*b*) their structural relation with one another. DS has focused mainly on lexical similarity relations, but semantic compositionality has attracted increasing attention in this area. In fact, DS

could not be regarded as a general model of meaning unless it were able to provide a satisfactory account of issues related to semantic compositionality.

The most common approach to compositionality in DS assumes that the distributional representation of a complex expression is a vector, and uses linear-algebraic operations to project lexical vectors to phrase vectors. The simplest form of vector combination is vector addition (Landauer & Dumais 1997) or some extension thereof (Kintsch 2001). Mitchell & Lapata (2010) compare several models based on vector addition and multiplication on a phrase similarity task, using ratings elicited from native speakers about adjective–noun, noun–noun, and verb–object combinations. Other models carry out composition in DS by representing lexemes and phrases with matrices and tensors, rather than with vectors alone (Zanzotto et al. 2010, Socher et al. 2012). In particular, the lexical function model aims to establish a direct link with Montague grammar by proposing a linear-algebraic representation of type-theoretic objects (Coecke et al. 2011, Baroni et al. 2014a, Grefenstette & Sadrzadeh 2015, Rimell et al. 2016). Arguments are vectors and functions taking arguments (e.g., verbs) are tensors; the number of arguments determines the order of tensors. Tensor-by-vector multiplication is the general composition method, as the distributional equivalent of function–argument application in formal semantics. Adopting Montague's idea that adjectives denote functions from the meaning of a noun onto the meaning of a modified noun, Baroni & Zamparelli (2010) build the vector **p** of an adjective–noun phrase like *interesting book* by multiplying a weight matrix **B** representing the adjective *interesting* by the noun vector **book**. Their model learns a separate matrix for each adjective with partial least-squares regression, using the dimensions of the vectors of the component nouns as independent variables and the adjective–noun vectors as the dependent variables.

The lexical function model is an attempt to represent formal semantic operations with DS; it produces interesting results when applied to adjective–noun modification (Vecchi et al. 2016, Asher et al. 2016), but has great difficulties in scaling up to multiargument sentences. Estimating the matrices and tensors for complex functional types such as transitive verbs can be very complex and may encounter data-sparseness problems. Paperno et al. (2014) propose a practical approximation of the lexical function model to address these limits, but it is hardly competitive with the much simpler additive models (Rimell et al. 2016). Vector addition is not fully adequate as a compositional operation, because it is commutative: If a sentence vector is the sum of its word vectors, *A dog bites a man* has exactly the same interpretation as *A man bites a dog*. Still, simple additive or multiplicative methods are very hard to beat by more complex distributional methods for semantic compositionality (Blacoe & Lapata 2012).

Representing complex expressions with vectors has the advantage of making lexical items and phrases directly comparable within the same vector space. The similarity between *carnivore* and the phrase *animal who eats meat* (Rimell et al. 2016) can be measured by computing the cosine between the respective vectors. Thus, the distributional approach to semantic similarity can be projected from the lexical level up to the level of the sentence and text similarity (Bentivogli et al. 2016). By contrast, it is doubtful whether representing the meaning of a sentence as a vector can capture complex inferences like those accounted for by symbolic representations. Sentence similarity is too shallow a task to test the adequacy of compositional semantic representations: Understanding the meaning of a sentence entails understanding a whole range of inferences that are licensed by it.

An alternative approach to compositional DS assumes that the representation of a sentence is not a vector but rather a logical form containing distributional vectors of the content words (Garrette et al. 2014, Asher et al. 2016, Beltagy et al. 2016). The aim is to exploit the complementary strengths of formal semantics and DS. The former has notorious difficulties in dealing with the richness, variability, vagueness, and gradience of lexical meaning (Boleda & Erk 2015, Boleda &

Herbelot 2016), but comes with a mathematically well-defined inferential system. Conversely, DS has very limited capabilities (if any) to account for inferences, but robust methods to learn fine-grained lexical representations. Formal structures can therefore provide the logical scaffolding to integrate lexical distributional representations into a full-fledged inferential model. Beltagy et al. (2016) present a system in which sentences are represented with first-order logical forms that are then used to perform inference with Markov logic, and DS is added to tackle near synonymy and lexical entailment. Chersoni et al. (2016) represent a sentence interpretation with a complex feature structure that formalizes a semantic frame including the distributional vectors of an event and its arguments. Semantic composition is modeled as an incremental process of feature unification to build the event represented by the sentence.

Research in compositional DS also deals with a broad range of phenomena that are especially challenging for formal semantics, such as meaning variation in context, selectional preferences, and coercion. Erk & Padó (2008) address the fact that, when words are composed, they tend to affect one another's meanings. The meaning of *run* in *The horse runs* is in fact different from its meaning in *The water runs* (Kintsch 2001). Erk & Padó (2008) claim that words are associated with various kinds of expectations (McRae et al. 2005), typical events for nouns and typical arguments for verbs that influence one another when words are combined, thereby altering their meaning. They model this context-sensitive compositionality by distinguishing the out-of-context vector of a word w_1 from its vector in the context of another word w_2. The vector-in-context for w_1 is obtained by combining (via addition or multiplication) the vector of w_1 with the vectors of the expectations activated by w_2. For instance, the vector-in-context assigned to *run* in *The horse runs* is obtained by combining the vector of *run* with the vectors of the most typical verbs in which *horse* appears as a subject (e.g., *gallop*, *trot*).

Selectional preferences are semantic constraints on the possible arguments of a predicate, and are usually represented with symbolic semantic types. Erk et al. (2010) propose a distributional model of selectional preferences in which the plausibility (i.e., thematic fit) of a noun n as an argument of a verb v is measured with the similarity in vector space between n and a set of noun exemplars occurring in the same argument role of v. Similarly, Baroni & Lenci (2010) measure the thematic fit of n by comparing its vector with a "prototype" vector obtained by averaging over the vectors of the most typical arguments of v. In both cases, the distributional measure of thematic fit is highly correlated with human ratings. Lenci (2011) has extended this model to account for the dynamic update of the semantic preferences of an argument, depending on how other roles in the sentence are filled. For example, given the agent *butcher*, the expected patient of the verb *cut* is likely to be *meat*, whereas given the agent *coiffeur*, the expected patient is likely to be *hair*. For other research on this topic, see Sayeed et al. (2016) and Tilk et al. (2016).

The same type of approach has also been used to model some cases of coercion, in which a predicate or argument meanings are adjusted to overcome a semantic preference violation (Pustejovsky 1995, Asher 2015). Chersoni et al. (2017) show that a verb object thematic fit computed with an incremental distributional model of sentence comprehension can reproduce the reading times of metonymic sentences like *The student began the book*, in which a type clash between an event-selecting verb and an entity-denoting object triggers the recovery of an implicit event, leading to extra processing costs. The same model can also identify the implicit event, accounting for its dependence on the verb subject (e.g., in *The student began the book*, the covert event is likely to be *read*, whereas in *The author began the book*, it is likely to be *write*). For a similar approach, see Zarcone et al. (2012, 2013).

The lexicon is often regarded as the bottleneck for formal semantics, but compositionality is surely the bottleneck for DS. How distributional representations can be projected from the lexical level to the sentence or even discourse level remains an open issue. Formal semantics models

rely on a clear definition of sentence meaning as truth conditions, but a clear understanding of a sentence distributional representation is lacking. The straightforward solution is to interpret sentence meaning as a vector, but doing so might not be sufficient to account for its inferential potential. A promising research avenue consists of looking for a "division of labor" between formal semantics and DS, representing sentences with logical forms enriched with distributional representations of lexical items (Beltagy et al. 2016, McNally 2017). Other interesting synergies could emerge from investigations of possible links between DS and dynamical formal semantic models (e.g., Kamp & Reyle 1993, Veltman 1996), which characterize the semantic content of linguistic expressions in terms of their context-change potential. The close connection between context and meaning is an important element of commonality between dynamic semantics and DS.

Most research on compositional DS has focused on providing a distributional model of classical Fregean compositionality. Still largely unexplored is the possibility of investigating alternative ways to build both compositional distributional representations, inspired by neuroscientific models of sentence comprehension (Baggio et al. 2012), and usage-based models of language, such as construction grammar (e.g., Goldberg 2006), which suggest that sentence meaning is built by dynamically activating and unifying semantic information associated with linguistic constructions. New perspectives can be achieved by investigating how to integrate DS, as a usage-based model of meaning, into the Constructionist framework in order to model behavioral data about sentence processing. For some preliminary results, see Chersoni et al. (2017) and Lebani & Lenci (2017).

6. CONCLUSIONS AND FUTURE CHALLENGES

DS is an active and lively research area in semantics, addressing a wide range of topics related to meaning. In addition to those analyzed in this review, further research issues in which DS is producing very interesting results include (*a*) the development of multimodal DSMs (Feng & Lapata 2010, Bruni et al. 2014) that integrate corpus-derived features and features extracted from images, which are also used to explore the interplay between linguistic and experiential information; (*b*) the study of polysemy, which uses DSMs to induce and represent difference senses from the distributional properties of lexical items (Schütze 1997, 1998, Heylen et al. 2015); and (*c*) the analysis of semantic change, which involves applying DS to diachronic corpora (Hamilton et al. 2016, Rodda et al. 2017) and investigating the change in productivity of syntactic constructions (Perek 2016). DS is a framework for semantic analysis that can provide new answers to classical semantic questions, as well as address problems that have often been ignored by other models of the lexicon.

DS builds semantic representations from co-occurrence statistics extracted from corpora as samples of language usage. In this way, DSMs are attuned to the large and multidimensional variability attested in language. By contrast, distributional representations are highly sensitive to the training corpus, as well as to the various parameters discussed above. Although this sensitivity is often regarded as a serious limitation of DS, it is consistent with the fact that semantic representations indeed depend strongly on context and vary across subjects and situations (Yee & Thompson-Schill 2016)—a feature that is often overlooked by traditional semantic models, which instead conceive lexical meaning as a largely static and invariant entity.

The real explanatory adequacy of DS as a model of meaning is far from clear. As discussed, DS often provides a quite coarse-grained representation of semantic content. Several aspects of meaning (e.g., quantification, intensionality, negation) are still unexplored and may lie beyond its scope. DS relies on continuous distributed representations, whose features are derived from corpus-based statistics. The limits of DS may arise from either of two factors: Some semantic

facts might not be handled in terms of nonsymbolic representations, and/or they might not have a correlate in distributional statistics harvested from corpus data. Both of these issues are worth exploring in future research.

As a model of the mental lexicon, DS has often been criticized because meaning cannot be reduced to co-occurrence statistics alone. At the same time, the importance of distributional data in forming semantic representations has been widely supported by empirical evidence. The contribution of linguistic experience vis-à-vis other kinds of nonlinguistic inputs in shaping concepts is an empirical question that is widely debated in cognitive science (Dove 2014, Louwerse 2011, Vigliocco et al. 2009). A fruitful perspective is to pursue a form of representational pluralism of meaning in which distributional statistics, extralinguistic experiential data, and symbolic aspects are integrated together in order to explain the richness of human semantic competence.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

I thank Magnus Sahlgren for useful discussions of many issues presented in this review.

LITERATURE CITED

Agirre E, Alfonseca E, Hall K, Kravalova J, Paca M, Soroa A. 2009. *A study on similarity and relatedness using distributional and WordNet-based approaches*. Presented at Conf. N. Am. Chapter Assoc. Comput. Linguist., Boulder, CO, May 31–June 5

Anderson AJ, Kiela D, Clark S, Poesio M. 2017. Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns. *Trans. ACL* 5:17–30

Asher N. 2015. Types, meanings and coercions in lexical semantics. *Lingua* 157:66–82

Asher N, Van de Cruys T, Bride A, Abrusán M. 2016. Integrating type theory and distributional semantics: a case study on adjective–noun compositions. *Comput. Linguist.* 42:703–25

Baggio G, van Lambalgen M, Hagoort P. 2012. The processing consequences of compositionality. In *The Oxford Handbook of Compositionality*, ed. M Werning, W Hinzen, E Machery, pp. 1–23. Oxford, UK: Oxford Univ. Press

Baroni M, Bernardi R, Do NQ, Chieh Shan C. 2012. Entailment above the word level in distributional semantics. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 23–32. Stroudsburg, PA: Assoc. Comput. Linguist.

Baroni M, Bernardi R, Zamparelli R. 2014a. Frege in space: a program of compositional distributional semantics. *Linguist. Issues Lang. Technol.* 9:5–110

Baroni M, Dinu G, Kruszewski G. 2014b. Don't count, predict! A systematic comparison of context-counting versus context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pp. 238–47. Stroudsburg, PA: Assoc. Comput. Linguist.

Baroni M, Lenci A. 2010. Distributional memory: a general framework for corpus-based semantics. *Comput. Linguist.* 36:673–721

Baroni M, Lenci A. 2011. How we BLESSed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics*, pp. 1–10. Stroudsburg, PA: Assoc. Comput. Linguist.

Baroni M, Zamparelli R. 2010. Nouns are vectors, adjectives are matrices: representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 1183–93. Stroudsburg, PA: Assoc. Comput. Linguist.

Beltagy I, Roller S, Cheng P, Erk K, Mooney RJ. 2016. Representing meaning with a combination of logical and distributional models. *Comput. Linguist.* 42:763–808

Bentivogli L, Bernardi R, Marelli M, Menini S, Baroni M, Zamparelli R. 2016. SICK through the SemEval glasses. Lesson learned from the evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. *Lang. Resour. Eval.* 50:95–124

Blacoe W, Lapata M. 2012. A comparison of vector-based representations for semantic composition. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Language Processing and Computational Natural Language Learning*, pp. 545–56. Stroudsburg, PA: Assoc. Comput. Linguist.

Blei D, Ng A, Jordan M. 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022

Boleda G, Erk K. 2015. Distributional semantic features as semantic primitives—or not. In *Proceedings of Knowledge Representation and Reasoning: Integrating Symbolic and Neural Approaches. Papers from the 2015 AAAI Spring Symposium*, pp. 2–5. Palo Alto, CA: AAAI

Boleda G, Herbelot A. 2016. Formal distributional semantics: introduction to the special issue. *Comput. Linguist.* 42:619–35

Bruni E, Tran NK, Baroni M. 2014. Multimodal distributional semantics. *J. Artif. Intell. Res.* 49:1–47

Budanitsky A, Hirst G. 2006. Evaluating WordNet-based measures of lexical semantic relatedness. *Comput. Linguist.* 32:13–47

Bullinaria J, Levy JP. 2007. Extracting semantic representations from word co-occurrence statistics: a computational study. *Behav. Res. Methods* 39:510–26

Bullinaria J, Levy JP. 2012. Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD. *Behav. Res. Methods* 44:890–907

Burgess C. 1998. From simple associations to the building blocks of language: modeling meaning in memory with the HAL model. *Behav. Res. Methods Instrum. Comput.* 30:188–98

Chersoni E, Blache P, Lenci A. 2016. Towards a distributional model of semantic complexity. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity*, pp. 12–22. Stroudsburg, PA: Assoc. Comput. Linguist.

Chersoni E, Blache P, Lenci A. 2017. Logical metonymy in a distributional model of sentence comprehension. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics*, pp. 168–77. Stroudsburg, PA: Assoc. Comput. Linguist.

Clark S. 2015. Vector space models of lexical meaning. In *Handbook of Contemporary Semantics*, ed. S Lappin, C Fox, pp. 493–522. Oxford, UK: Wiley-Blackwell. 2nd ed.

Coecke B, Sadrzadeh M, Clark S. 2011. Mathematical foundations for a compositional distributional model of meaning. *Linguist. Anal.* 36:345–84

Collobert R, Weston J. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 160–67. New York: ACM

Curran JR. 2003. *From distributional to semantic similarity*. PhD thesis, Univ. Edinburgh, Edinburgh, UK

Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. 1990. Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.* 41:391–407

Dove G. 2014. Thinking in words: language as an embodied medium of thought. *Top. Cogn. Sci.* 6:371–89

Elman JL. 1990. Finding structure in time. *Cogn. Sci.* 14:179–211

Erk K. 2012. Vector space models of word meaning and phrase meaning: a survey. *Linguist. Lang. Compass* 6:635–53

Erk K. 2016. What do you know about an alligator when you know the company it keeps? *Semant. Pragmat.* 9:1–63

Erk K, Padó S. 2008. A structured vector space model for word meaning in context. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 897–906. Stroudsburg, PA: Assoc. Comput. Linguist.

Erk K, Padó S, Padó U. 2010. A flexible, corpus-driven model of regular and inverse selectional preferences. *Comput. Linguist.* 36:723–63

Evert S. 2008. Corpora and collocations. In *Corpus Linguistics: An International Handbook*, ed. A Lüdeling, M Kytö, pp. 1212–48. Berlin: de Gruyter

Feng Y, Lapata M. 2010. Visual information in semantic representation. In *Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 91–99. Stroudsburg, PA: Assoc. Comput. Linguist.

Finkelstein L, Gabrilovich E, Matias Y, Rivlin E, Solan Z, et al. 2001. Placing search in context: the concept revisited. *ACM Trans. Inf. Syst.* 20:116–31

Firth JR. 1957. A synopsis of linguistic theory, 1930–1955. In *Studies in Linguistic Analysis*, pp. 1–32. Oxford, UK: Blackwell

Garrette D, Erk K, Mooney RJ. 2014. A formal approach to linking logical form and vector-space lexical semantics. In *Computing Meaning*, ed. H Bunt, J Bos, S Pulman, 4:27–47. Dordrecht, Neth.: Springer

Garvin PL. 1962. Computer participation in linguistic research. *Language* 38:385–89

Gerz D, Vuli I, Hill F, Reichart R, Korhonen A. 2016. SimVerb-3500: a large-scale evaluation set of verb similarity. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2173–82. Stroudsburg, PA: Assoc. Comput. Linguist.

Goldberg AE. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford, UK: Oxford Univ. Press

Grefenstette E, Sadrzadeh M. 2015. Concrete models and empirical evaluations for the categorical compositional distributional model of meaning. *Comput. Linguist.* 41:71–118

Griffiths T, Tenenbaum J, Steyvers M. 2007. Topics in semantic representation. *Psychol. Rev.* 114:211–44

Hamilton WL, Leskovec J, Jurafsky D. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1489–510. Stroudsburg, PA: Assoc. Comput. Linguist.

Harris ZS. 1954. Distributional structure. *Word* 10:146–62

Heylen K, Wielfaert T, Speelman D, Geeraerts D. 2015. Monitoring polysemy: word space models as a tool for large-scale lexical semantic analysis. *Lingua* 157:153–72

Hill F, Reichart R, Korhonen A. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Comput. Linguist.* 41:665–95

Jones MN, Kintsch W, Mewhort DJK. 2006. High-dimensional semantic space accounts of priming. *J. Mem. Lang.* 55:534–52

Jones MN, Mewhort DJK. 2007. Representing word meaning and order information in a composite holographic lexicon. *Psychol. Rev.* 22:701–8

Jones MN, Willits JA, Dennis S. 2015. Models of semantic memory. In *Oxford Handbook of Mathematical and Computational Psychology*, ed. JR Busemeyer, Z Wang, JT Townsend, A Eidels, pp. 232–54. Oxford, UK: Oxford Univ. Press

Jurafsky D, Martin JA. 2008. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ: Prentice Hall. 2nd ed.

Kamp H, Reyle U. 1993. *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Dordrecht, Neth.: Kluwer

Kanerva P, Kristofersson J, Holst A. 2000. Random indexing of text samples for latent semantic analysis. In *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society*, p. 1036. Mahwah, NJ: Erlbaum

Keuleers E, Balota DA. 2015. Megastudies, crowdsourcing, and large datasets in psycholinguistics: an overview of recent developments. *Q. J. Exp. Psychol.* 68:1457–68

Kiela D, Clark S. 2014. A systematic study of semantic vector space model parameters. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and Their Compositionality*, pp. 21–30. Stroudsburg, PA: Assoc. Comput. Linguist.

Kintsch W. 2001. Predication. *Cogn. Sci.* 25:173–202

Kotlerman L, Dagan I, Szpektor I, Zhitomirsky-Geffet M. 2010. Directional distributional similarity for lexical inference. *Nat. Lang. Eng.* 16:359–89

Landauer TK, Dumais S. 1997. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.* 104:211–40

Lapesa G, Evert S. 2014. A large scale evaluation of distributional semantic models: parameters, interactions and model selection. *Trans. ACL* 2:531–45

Lapesa G, Evert S. 2017. Large-scale evaluation of dependency-based DSMs: Are they worth the effort? In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 394–400. Stroudsburg, PA: Assoc. Comput. Linguist.

Lebani GE, Lenci A. 2017. Modelling the meaning of argument constructions with distributional semantics. In *Proceedings of the AAAI 2017 Spring Symposium on Computational Construction Grammar and Natural Language Understanding*, pp. 197–204. Palo Alto, CA: AAAI

Lenci A. 2008. Distributional approaches in linguistic and cognitive research. *Ital. J. Linguist.* 20:1–31

Lenci A. 2011. Composing and updating verb argument expectations: a distributional semantic model. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pp. 58–66. Stroudsburg, PA: Assoc. Comput. Linguist.

Lenci A, Benotto G. 2012. Identifying hypernyms in distributional semantic spaces. In *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics*, pp. 75–79. Stroudsburg, PA: Assoc. Comput. Linguist.

Levy O, Goldberg Y. 2014a. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pp. 302–8. Stroudsburg, PA: Assoc. Comput. Linguist.

Levy O, Goldberg Y. 2014b. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the 18th Conference on Computational Language Learning*, pp. 171–80. Stroudsburg, PA: Assoc. Comput. Linguist.

Levy O, Goldberg Y, Dagan I. 2015a. Improving distributional similarity with lessons learned from word embeddings. *Trans. ACL* 3:211–25

Levy O, Remus S, Biemann C, Dagan I. 2015b. Do supervised distributional methods really learn lexical inference relations? In *Proceedings of the 2015 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 970–76. Stroudsburg, PA: Assoc. Comput. Linguist.

Lofi C. 2015. Measuring semantic similarity and relatedness with distributional and knowledge-based approaches. *Inf. Media Technol.* 10:493–501

Louwerse MM. 2011. Symbol interdependency in symbolic and embodied cognition. *Top. Cogn. Sci.* 3:273–302

Mandera P, Keuleers E, Brysbaert M. 2015. How useful are corpus-based methods for extrapolating psycholinguistic variables? *Q. J. Exp. Psychol.* 68:1623–42

Mandera P, Keuleers E, Brysbaert M. 2017. Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: a review and empirical validation. *J. Mem. Lang.* 92:57–78

Manning C, Schütze H. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press

McNally L. 2017. Kinds, descriptions of kinds, concepts, and distributions. In *Bridging Formal and Conceptual Semantics: Selected Papers of BRIDGE-14*, ed. K Balogh, W Petersen, pp. 39–61. Düsseldorf, Ger.: Düsseldorf Univ. Press

McRae K, Hare M, Elman JL, Ferretti TR. 2005. A basis for generating expectancies for verbs from nouns. *Mem. Cogn.* 33:1174–84

McRae K, Jones MN. 2013. Semantic memory. In *The Oxford Handbook of Cognitive Psychology*, ed. D Reisberg, pp. 206–19. Oxford, UK: Oxford Univ. Press

Mikolov T, Chen K, Corrado GS, Dean J. 2013a. Efficient estimation of word representations in vector space. arXiv:1301.3781 [cs.CL]

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th Conference on Advances in Neural Information Processing Systems*, pp. 3111–19. **https://papers.nips.cc/book/advances-in-neural-information-processing-systems-26-2013**

Mikolov T, Yih W-T, Zweig G. 2013c. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 746–51. Stroudsburg, PA: Assoc. Comput. Linguist.

Miller G, Charles W. 1991. Contextual correlates of semantic similarity. *Lang. Cogn. Process.* 6:1–28

Miller GA. 1967. Empirical methods in the study of semantics. In *Journeys in Science: Small Steps—Great Strides*. Albuquerque: Univ. N. M. Press

Mitchell J, Lapata M. 2010. Composition in distributional models of semantics. *Cogn. Sci.* 34:1388–429
Mitchell TM, Shinkareva SV, Carlson A, Chang K-M, Malave VL, et al. 2008. Predicting human brain activity associated with the meanings of nouns. *Science* 320:1191–95
Mohammad SM, Dorr BJ, Hirst G, Turney PD. 2013. Computing lexical contrast. *Comput. Linguist.* 39:1–60
Morris J, Hirst G. 2004. Non-classical lexical semantic relations. In *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics*, pp. 46–51. Stroudsburg, PA: Assoc. Comput. Linguist.
Murphy ML. 2003. *Semantic Relations and the Lexicon: Antonymy, Synonymy, and the Other Paradigms*. Cambridge, UK: Cambridge Univ. Press
Osgood CE. 1952. The nature and measurement of meaning. *Psychol. Bull.* 49:197–237
Padó S, Lapata M. 2007. Dependency-based construction of semantic space models. *Comput. Linguist.* 33:161–99
Paperno D, Pham NT, Baroni M. 2014. A practical and linguistically-motivated approach to compositional distributional semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pp. 90–99. Stroudsburg, PA: Assoc. Comput. Linguist.
Pennington J, Socher R, Manning CD. 2014. GloVe: global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1532–43. Stroudsburg, PA: Assoc. Comput. Linguist.
Perek F. 2016. Using distributional semantics to study syntactic productivity in diachrony: a case study. *Linguistics* 54:149–88
Pustejovsky J. 1995. *The Generative Lexicon*. Cambridge, MA: MIT Press
Recchia G, Jones MN, Sahlgren M, Kanerva P. 2015. Encoding sequential information in vector space models of semantics: comparing holographic reduced representation and random permutation. *Comput. Intell. Neurosci.* 2015:986574
Rimell L. 2014. Distributional lexical entailment by topic coherence. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 511–19. Stroudsburg, PA: Assoc. Comput. Linguist.
Rimell L, Maillard J, Polajnar T, Clark S. 2016. RELPRON: a relative clause evaluation data set for compositional distributional semantics. *Comput. Linguist.* 42:661–701
Riordan B, Jones MN. 2011. Redundancy in perceptual and linguistic experience: comparing feature-based and distributional models of semantic representation. *Top. Cogn. Sci.* 3:303–45
Rodda M, Senaldi MSG, Lenci A. 2017. *Panta rei*: tracking semantic change with distributional semantics in Ancient Greek. *Ital. J. Comput. Linguist.* 3:11–24
Roller S, Erk K, Boleda G. 2014. Inclusive yet selective: supervised distributional hypernymy detection. In *Proceedings of the 25th International Conference on Computational Linguistics*, pp. 1025–36. Stroudsburg, PA: Assoc. Comput. Linguist.
Rubenstein H, Goodenough JB. 1965. Contextual correlates of synonymy. *Commun. ACM* 8:627–33
Sahlgren M. 2006. *The word-space model. Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.* PhD thesis, Stockholm Univ., Stockholm, Swed.
Sahlgren M, Lenci A. 2016. The effects of data size and frequency range on distributional semantic models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 975–80. Stroudsburg, PA: Assoc. Comput. Linguist.
Salton G, Wong A, Yang CS. 1975. A vector space model for automatic indexing. *Commun. ACM* 18:613–20
Santus E, Lenci A, Lu Q, Schulte im Walde S. 2014a. Chasing hypernyms in vector spaces with entropy. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 38–42. Stroudsburg, PA: Assoc. Comput. Linguist.
Santus E, Lu Q, Lenci A, Huang CR. 2014b. Taking antonymy mask off in vector space. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*, pp. 135–44. Stroudsburg, PA: Assoc. Comput. Linguist.
Santus E, Yung F, Lenci A, Huang CR. 2015. EVALution 1.0: an evolving semantic dataset for training and evaluation of distributional semantic models. In *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications*, pp. 64–69. Stroudsburg, PA: Assoc. Comput. Linguist.

Sayeed A, Greenberg C, Demberg V. 2016. Thematic fit evaluation: an aspect of selectional preferences. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for Natural Language Processing*, pp. 99–105. Stroudsburg, PA: Assoc. Comput. Linguist.

Schütze H. 1997. *Ambiguity Resolution in Language Learning: Computational and Cognitive Models*. Stanford, CA: Cent. Study Lang. Inf.

Schütze H. 1998. Automatic word sense discrimination. *Comput. Linguist.* 24:97–123

Shaoul C, Westbury CF. 2010. Exploring lexical co-occurrence space using HiDEx. *Behav. Res. Methods* 42:393–413

Shwartz V, Santus E, Schlechtweg D. 2017. Hypernyms under siege: linguistically-motivated artillery for hypernymy detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 65–75. Stroudsburg, PA: Assoc. Comput. Linguist.

Socher R, Huval B, Manning CD, Ng AY. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 1201–11. Stroudsburg, PA: Assoc. Comput. Linguist.

Tilk O, Demberg V, Sayeed AB, Klakow D, Thater S. 2016. Event participant modelling with neural networks. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 171–82. Stroudsburg, PA: Assoc. Comput. Linguist.

Turney PD. 2006. Similarity of semantic relations. *Comput. Linguist.* 32:379–416

Turney PD. 2008. A uniform approach to analogies, synonyms, antonyms, and associations. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pp. 905–12. Stroudsburg, PA: Assoc. Comput. Linguist.

Turney PD, Pantel P. 2010. From frequency to meaning: vector space models of semantics. *J. Artif. Intell. Res.* 37:141–88

Van de Cruys T. 2008. A comparison of bag of words and syntax-based approaches for word categorization. In *Proceedings of the Lexical Semantics Workshop: Bridging the Gap Between Semantic Theory and Computational Simulations*, pp. 47–54. Amsterdam: Univ. Amsterdam

Vecchi EM, Marelli M, Zamparelli R, Baroni M. 2016. Spicy adjectives and nominal donkeys: capturing semantic deviance using compositionality in distributional spaces. *Cogn. Sci.* 41:102–36

Veltman F. 1996. Defaults in update semantics. *J. Philos. Logic* 25:221–61

Vigliocco G, Meteyard L, Andrews M, Kousta ST. 2009. Toward a theory of semantic representation. *Lang. Cogn.* 1:219–47

Weeds J, Clarke D, Reffin J, Weir D, Keller B. 2014. Learning to distinguish hypernyms and co-hyponyms. In *Proceedings of the 25th International Conference on Computational Linguistics*, pp. 2249–59. Stroudsburg, PA: Assoc. Comput. Linguist.

Weeds J, Weir D, McCarthy D. 2004. Characterising measures of lexical distributional similarity. In *Proceedings of the 20th International Conference on Computational Linguistics*, pp. 1–7. Stroudsburg, PA: Assoc. Comput. Linguist.

Wittgenstein L. 1953. *Philosophical Investigations*, transl. GEM Anscombe. Oxford, UK: Blackwell

Yee E, Thompson-Schill SL. 2016. Putting concepts into context. *Psychon. Bull. Rev.* 23:1015–27

Yih W, Zweig G, Platt JC. 2012. Polarity inducing latent semantic analysis. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 1212–22. Stroudsburg, PA: Assoc. Comput. Linguist.

Zanzotto FM, Korkontzelos I, Fallucchi F, Manandhar S. 2010. Estimating linear models for compositional distributional semantics. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 1263–71. Stroudsburg, PA: Assoc. Comput. Linguist.

Zarcone A, Lenci A, Padó S, Utt J. 2013. Fitting, not clashing! A distributional semantic model of logical metonymy. In *Proceedings of the 10th International Conference on Computational Semantics*, pp. 404–10. Stroudsburg, PA: Assoc. Comput. Linguist.

Zarcone A, Utt J, Padó S. 2012. Modeling covert event retrieval in logical metonymy: probabilistic and distributional accounts. In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics*, pp. 70–79. Stroudsburg, PA: Assoc. Comput. Linguist.