ANNUAL REVIEWS

Annual Review of Linguistics

Distributional Semantics and Linguistic Theory

Gemma Boleda[1,2]

[1]Department of Translation and Language Sciences, Universitat Pompeu Fabra, Barcelona 08018, Spain; email: gemma.boleda@upf.edu

[2]Catalan Institution for Research and Advanced Studies (ICREA), Barcelona 08010, Spain

Annu. Rev. Linguist. 2020. 6:213–34

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

https://doi.org/10.1146/annurev-linguistics-011619-030303

Keywords

distributional semantics, vector space models, vector spaces, semantic spaces, computational semantics, semantic change, diachronic semantics, polysemy, composition, syntax–semantics interface, derivational morphology

Abstract

Distributional semantics provides multidimensional, graded, empirically induced word representations that successfully capture many aspects of meaning in natural languages, as shown by a large body of research in computational linguistics; yet, its impact in theoretical linguistics has so far been limited. This review provides a critical discussion of the literature on distributional semantics, with an emphasis on methods and results that are relevant for theoretical linguistics, in three areas: semantic change, polysemy and composition, and the grammar–semantics interface (specifically, the interface of semantics with syntax and with derivational morphology). The goal of this review is to foster greater cross-fertilization of theoretical and computational approaches to language as a means to advance our collective knowledge of how it works.

1. INTRODUCTION

This review provides a critical discussion of the literature on distributional semantics, with an emphasis on methods and results that are relevant for theoretical linguistics, in three areas: semantic change, polysemy and composition, and the grammar–semantics interface. Distributional semantics has proven useful in computational linguistics and cognitive science (Landauer & Dumais 1997, Schütze 1992; see Clark 2015); yet, its impact in theoretical linguistics has so far been limited. Greater cross-fertilization of theoretical and computational approaches promises to advance our knowledge of how language works, and fostering such cross-fertilization is the ultimate goal of this review. Accordingly, I cover mostly research within computational linguistics, rather than cognitive science.

1.1. Distributional Semantics in a Nutshell

In this section, I provide only a brief introduction to distributional semantics, such that the review is self-contained (for more comprehensive introductions, see Clark 2015, Erk 2012, and Lenci 2018). Distributional semantics is based on the Distributional Hypothesis, which states that similarity in meaning results in similarity of linguistic distribution (Harris 1954). Words that are semantically related, such as *postdoc* and *student*, are used in similar contexts, as in *a poor _, the _ struggled through the deadline* (Boleda & Herbelot 2016, p. 623). Distributional semantics reverse engineers the process and induces semantic representations from contexts of use.

In its most basic and frequent form (**Figure 1**), distributional semantics represents word meaning by taking large amounts of text as input and, through an abstraction mechanism, producing a distributional model, akin to a lexicon, with semantic representations in the form of vectors—essentially, lists of numbers that determine points in a multidimensional space. However, many more possibilities are available and have been experimented with. The definition of distributional semantics encompasses all kinds of contexts, including, for instance, the visual context in

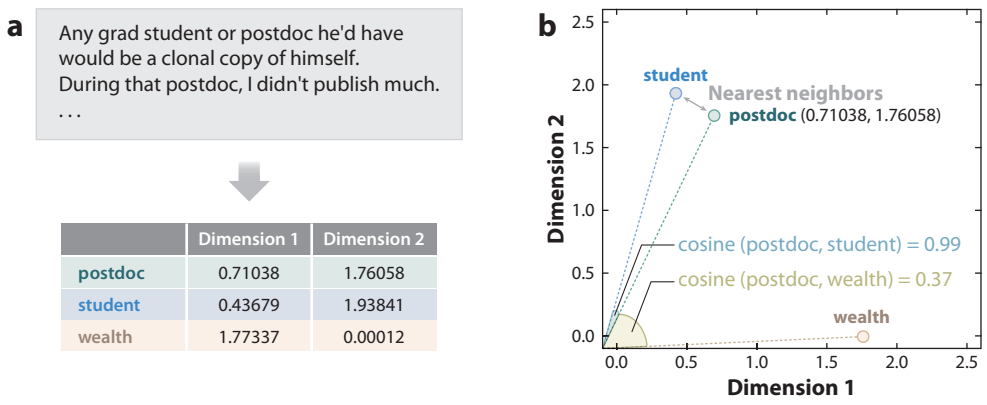

Figure 1

Distributional semantics in a nutshell. (*a*) Inducing semantic representations from natural language data. The arrow represents the abstraction mechanism that produces a distributional model from the text. (*b*) Visualizing and operating with distributional representations. Words are points in a space determined by the values in the dimensions of their vectors, such as 0.71038 and 1.76058 for *postdoc*. *Postdoc* and *student* are nearer in semantic space than are *postdoc* and *wealth*, and in fact they are nearest neighbors of (i.e., words closest to) one another. Figure adapted from Boleda & Herbelot (2016, figure 1) under a Creative Commons Attribution (CC BY) 3.0/4.0 License.

which words are used (Baroni 2016b); some models take morphemes, phrases, sentences, or documents instead of words as units to represent (Turney & Pantel 2010); and units can be represented via more complex algebraic objects than vectors, such as matrices or tensors (Grefenstette & Sadrzadeh 2011).

The collection of units in a distributional model constitutes a vector space or semantic space, in which semantic relations can be modeled as geometric relations. Vectors determine points in space; **Figure 1*b*** is a graphical rendering of our toy lexicon. The vectors for *postdoc* and *student* are closer together in space than are those of *postdoc* and *wealth*, because their vector values are more similar. The abstraction mechanisms used to obtain distributional models are such that similar contexts of use result in similar vectors; therefore, vector similarity correlates with distributional similarity, which in turn correlates with semantic similarity or, more generally, semantic relatedness. The most common similarity measure in distributional semantics is the cosine of the angle between two vectors: The closer the vectors are to one another, the larger the cosine similarity is. For instance, the cosine between *postdoc* and *student* in our space is 0.99, whereas it is 0.37 for *postdoc* versus *wealth* (cosine values for positive vectors range between 0 and 1).

Our example is two-dimensional, but in actual distributional models many more dimensions (frequently 300–400) are used. While we cannot represent so many dimensions visually, the geometric properties of two-dimensional spaces discussed here apply to any number of dimensions. Given that real distributional vectors are not directly interpretable, a very common way for researchers to gain insight into the information encoded in word vectors is to inspect their nearest neighbors. These are the words that are closest to a given target; for instance, *student* is the nearest neighbor of *postdoc* in the mini semantic space.

Finally, there are many different versions of the abstraction function (**Figure 1**). Earlier distributional models were built by extracting and transforming co-occurrence statistics, while models based on neural networks have recently gained ground due to their good performance (Baroni et al. 2014b). Neural networks are a versatile machine-learning type of algorithm, used for tasks such as machine translation or image labeling. For reasons of scope, in this review I cover only uses of neural networks that are specifically targeted at building semantic spaces akin to those in classic distributional semantics.

1.2. Distributional Semantics as a Model of Word Meaning

Distributional semantics arises largely from structuralist traditions (Sahlgren 2008). As in structuralism, words are defined according to their position in a system, the lexicon, on the basis of a set of features; their values are defined by contrasts in the words' contexts of use. However, in structuralism usually only a few features are used; they are defined manually; and they have an intrinsic meaning. For instance, they can be semantic primitives of the sort ±MALE. As Boleda & Erk (2015) point out, in distributional semantics the individual features lack an intrinsic meaning, and what gains prominence are the geometric relationships between the words. Semantic notions such as ±MALE are instead captured in a distributed fashion, as varying patterns across the whole vector. There are three further key differences from traditional feature-based approaches in linguistics that render distributional semantics attractive as a model of word meaning.

The first is the fact that distributional representations are learned from natural language data, and thus are radically empirical. The induction process is automatic, scaling up to very large vocabularies and any language or domain with enough linguistic data to process. For instance, Bojanowski et al. (2017) provide semantic spaces for 157 languages, built from Wikipedia text. Distributional semantics thus provides semantic representations on a large scale, in a single, coherent system in which systematic explorations are possible.

Table 1 Near-synonyms in semantic space: the words closest to *man*, *chap*, *lad*, and *guy*

Word	Nearest neighbors[a]
man	woman, gentleman, gray-haired, boy, person
lad	boy, bloke, scouser, lass, youngster
chap	bloke, guy, lad, fella, man
guy	bloke, chap, doofus, dude, fella

[a]Nearest neighbors are from the distributional model of Baroni et al. (2014b).
Table adapted from Baroni (2016a).

The second difference is high multidimensionality. The information abstracted from the data is distributed across all the dimensions of a vector, typically a few hundred, allowing rich and nuanced information to be encoded. In traditional approaches, again for methodological and practical reasons, comparatively few features are specified. Semantic distinctions can be very subtle, as shown by the phenomenon of near-synonymy. All the words in **Table 1** (*man*, *chap*, *lad*, and *guy*) denote male adult humans, but each presents different nuances that are difficult to express in a symbolic system using few features. Their nearest neighbors illustrate the capacity of distributional semantic models to capture both generic and specific semantic features. On the one hand, most of the neighbors are human- or male-denoting words, suggesting that information akin to semantic features in decompositional approaches, such as ±MALE, is captured in the space (Mikolov et al. 2013b provide quantitative evidence). On the other hand, the nearest neighbors reflect semantic differences between them, such as *lad* being used for younger men (its closest word in the space is *boy*, and it is also near *lass*, used to refer to girls in some English dialects, and *youngster*).

The third, and related, difference is gradedness. The information in the vectors is expressed in the form of continuous values, and measures such as cosine similarity are graded. Two vectors can be more or less similar, or similar in certain dimensions but not others. In the example in **Table 1**, even if all four words are near-synonyms, *chap* and *guy* are nearer near-synonyms, if we go by the standard test for synonymy in linguistics (substitutability in context; Lyons 1977). Correspondingly, their vectors are the closest of the set, as shown by their sharing many nearest neighbors.

2. SEMANTIC CHANGE

In diachronic semantics, especially lexical semantic change, the interaction between use and meaning (crucial for distributional semantics) has traditionally been the focus of interest for theoretical linguists (Deo 2015, Traugott & Dasher 2001). For instance, the contexts of use of *gay* reflect a gradual change in meaning during the twentieth century: from a meaning similar to 'cheerful' to its current predominant use as 'homosexual.' This can be seen in the contrast between the sentences in example 1, which are from the year 1900, and the sentences in example 2, which are from 2000 (Davies 2010):

(1) She was a fine-looking woman, cheerful and gay.
We assembled around the breakfast with spirits as gay and appetites as sharp as ever.

(2) […] the expectation that effeminate men and masculine women are more likely to be seen as gay men and lesbians, respectively.
'I don't personally support gay marriage myself,' Edwards said.

The three key properties of distributional semantics mentioned above are useful to model semantic change, and this is currently a blooming topic in computational linguistics (for overviews, see

Kutuzov et al. 2018, Tahmasebi et al. 2018). High dimensionality allows distributional semantics to represent many semantic nuances that can be subject to change; gradedness in representations is crucial to account for the gradual nature of change; and, as discussed below, its data-driven nature allows it to induce semantic change from changes in usage.

2.1. Distributional Approaches to Diachronic Semantics

Scholars began using distributional methods for semantic change around the 2010s, with earlier research (Gulordava & Baroni 2011, Sagi et al. 2009) using classic distributional methods and Kim et al. (2014) introducing neural network representations, which have been predominant in later work (Del Tredici et al. 2019, Hamilton et al. 2016, Szymanski 2017). Distributional approaches are based on the hypothesis that a change in context of use mirrors a change in meaning, which can be regarded as a special case of the Distributional Hypothesis. Thus, these approaches infer a change in meaning when they observe a change in the context of use.

This inferential process is typically carried out by building word representations at different points in time and comparing them [although Rosenfeld & Erk (2018) include time as a variable in the model instead]. This method is used to both detect semantic change and track its temporal evolution. For instance, Kim et al. (2014) built one distributional lexicon per year from 1850 to 2009 using data from the Google Books Ngrams corpus (Michel et al. 2011). The cosine similarity of the word *gay*, when compared with its representation in the 1900 lexicon, goes down during the twentieth century, with the drop accelerating from around 0.75 at the end of the 1970s to around 0.3 in 2000.

Figure 2 depicts the trajectory of three words across time in another study (Hamilton et al. 2016). It illustrates how inspection of nearest neighbors can help trace the specific meaning shift taking place. In 1900, *gay* is near words such as *daft* or *cheerful*, and by 1990, it is near *homosexual*. The change in *broadcast*, from a concrete to a more abstract meaning, is metaphorical in nature (from spreading seeds to spreading information or signal), and *awful* undergoes pejoration, from a positive to a negative denotation. Another method used to track specific semantic changes involves targeted comparisons to words related to the old and the new meanings. For instance, Kim et al. (2014) compare the evolution of the cosine similarities of *cell* to *dungeon* and *phone* over the years. However, the latter approach requires previous knowledge of the specific change taking place.

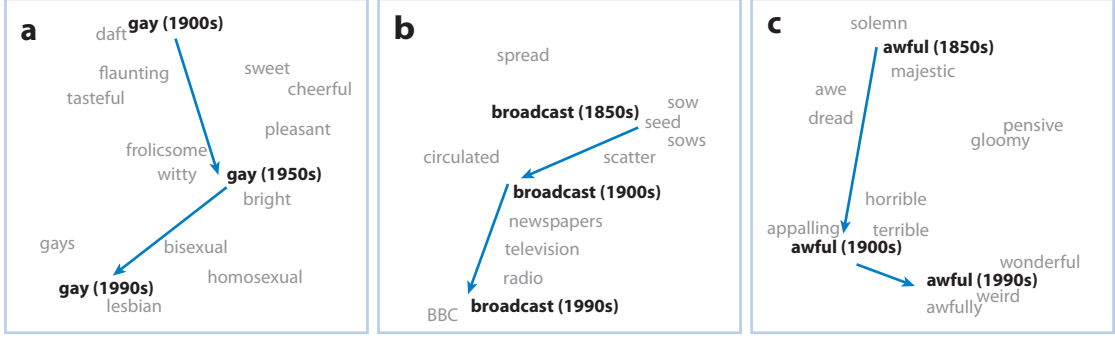

Figure 2

Two-dimensional visualization of semantic change for three English words across the 19th and 20th centuries. Nearest neighbors are shown in gray font along the words of interest. This two-dimensional figure was obtained via dimensionality reduction from a space with 300-dimensional vectors. Figure adapted with permission from Hamilton et al. (2016).

Current experiments are addressing two related areas (Tahmasebi et al. 2018): sense-specific semantic change, in which sense representations are induced and then tracked (also see Section 3.2), and detection of not only the presence but also the type of semantic shift. The latter literature, starting with pioneering research by Sagi et al. (2009), presents some evidence that distributional methods can spot narrowing and broadening, two classically described types of diachronic shift (Hock 1991). A case of narrowing is 'deer,' which evolved from Old English *deor*, meaning 'animal,' to its current, narrower denotation; an example of broadening is *dog*, from Late Old English *docga*, which used to denote a specific breed of dog, to its current, broader meaning. An extreme form of broadening results in grammaticalization, as in the verb *do* going from a lexical to an auxiliary verb between the fifteenth and eighteenth centuries. Sagi et al. (2009) trace these three words by representing each context of use individually, with one vector per sentence. They show that contexts become more separate over time for *dog* and *do*, corresponding to the broadening effect, and that the reverse occurs for *deer*. Moreover, their distributional measure correlates with the proportion of periphrastic uses of *do* through the centuries, independently estimated via manual annotation of texts.

To date, most research has focused on showing that distributional semantics can model semantic change, rather than on systematically exploring data and advancing our knowledge of the phenomenon. An exception is a study by Xu & Kemp (2015), who assessed two previously proposed laws that make contradicting predictions. Their large-scale computational analysis, based on distributional semantic models of English on the Google Books Ngrams corpus, shows that pairs of synonyms tend to stay closer together in space than do control pairs across the twentieth century, in four data sets jointly comprising tens of thousands of words. Thus, these authors provide support for the law of parallel change (Stern 1921), which posits that related words undergo similar changes, and against the law of differentiation (Bréal 1897), which defends the idea that synonyms tend to evolve different meanings because it is not efficient for languages to maintain synonyms. Other generalizations about semantic change emerging from research with distributional methods have been proposed, but controlled experiments have called them into question (Dubossarsky et al. 2017).

2.2. Discussion

Distributional semantics has tremendous potential to accelerate research in semantic change, in particular, the exploration of large-scale diachronic data, in four main crucial ways: (*a*) detecting semantic change, as a change in the representation of a word across time; (*b*) temporally locating it, by monitoring the rate of change in the distributional representation; (*c*) tracking the specific semantic evolution of the word, via an inspection of the nearest neighbors or targeted examination of cosine similarities; and (*d*) testing competing theories, via large-scale empirical studies. It can also help detect the type of semantic change, although this is still an underresearched topic.

A major challenge is that distributional methods, especially those based on neural networks, are quite data hungry, while many data sets in diachronic semantics are rather small (Kutuzov et al. 2018). This means that most studies are for English, and other languages are neglected: Of 23 data sets used for diachronic semantics, identified in a survey by Tahmasebi et al. (2018), only 4 are not in English. Moreover, the vast majority of studies focus on the Google Books Ngrams corpus, which covers only the period between 1850 and 2009.

When the data are scarce, spurious effects easily arise. For instance, in their study of meaning shift in a community of soccer fans, using data from 2011 to 2017, Del Tredici et al. (2019) find that reference to specific people or events causes changes in cosine similarity that do not correspond to semantic change; an example is *stubborn*, which in 2017 was used mostly when talking about a

new coach.[1] Such effects challenge the Distributional Hypothesis, as a change in context does not signal a change in meaning, and call for more nuanced methods. This kind of issue is typically less problematic for studies involving longer timescales, because of the larger quantity and variety of data, but it can arise when data are scarce or when there are systematic differences in the sources for different time periods—for instance, if texts are from different genres.

Another issue is that research has focused mostly on lexical semantic change, whereas in diachronic semantics there has been much work on grammaticalization processes (Deo 2015). While classic distributional approaches could not account for function words (to the point that they were typically removed from the vocabulary), recent neural network models provide usable representations for them (Mikolov et al. 2010, Peters et al. 2018), leading to new possibilities.

3. POLYSEMY AND COMPOSITION

Words are notoriously ambiguous or polysemous; that is, they adopt different meanings in different contexts (see Cruse 1986, among many others). For instance, *postdoc* refers to a person in the first sentence shown in **Figure 1*a***, and to a period of time in the second. Distributional semantics has traditionally addressed this issue in two ways, which resonate with linguistic treatments of polysemy (Lyons 1977). The first, and by far the most predominant, approach is to take the word as a unit of representation and provide a single representation that encompasses all of its uses (Section 3.1). The second approach is to provide different vectors for different word senses (Section 3.2).

3.1. Single Representation, Polysemy via Composition

The predominant, single-representation approach is similar in spirit to structured approaches to the lexicon, such as the Generative Lexicon (Pustejovsky 1995), Frame Semantics (Fillmore 2006), or Head-Driven Phrase Structure Grammar (Pollard & Sag 1994), even if not directly inspired by them. These approaches aim to encode all the relevant information in the lexical entry, and then define mechanisms to deploy the right meaning in context, usually by composition. As an example, Pustejovsky (1995, pp. 122–23) formalizes two readings of *bake*, a change of state (*John baked the potato*) and a creation sense (*John baked the cake*), by letting the lexical entries of the verb and the noun interact. If *bake* combines with a mass-denoting noun, then the change of state sense emerges; if it combines with an artifact, then the creation sense emerges. This approach has the advantage of capturing aspects of meaning that are common to the different contexts, while being able to account for the differences. Sense accounts of polysemy struggle with this, and face a host of other serious theoretical, methodological, and empirical issues (for discussion, see Kilgarriff 1997, Pustejovsky 1995).

In standard distributional semantics, each word is assigned a single vector, which is an abstraction over all its contexts of use, thus encompassing all the word senses that are attested in the data (Arora et al. 2018). Pioneering research by Kintsch (2001) in cognitive science extended distributional methods to composing such generic word representations into larger constituents. The computational linguistics community took up this line of research almost a decade later (Baroni & Zamparelli 2010, Coecke et al. 2011, Erk & Padó 2008, Mikolov et al. 2013a, Mitchell & Lapata 2010, Socher et al. 2012). Compositional distributional methods build representations of phrases out of the representations of their parts, and the corresponding meaning is expected to emerge

[1]Del Tredici et al. (2019) detect such cases with a complementary distributional measure, based on the specificity of the contexts of use.

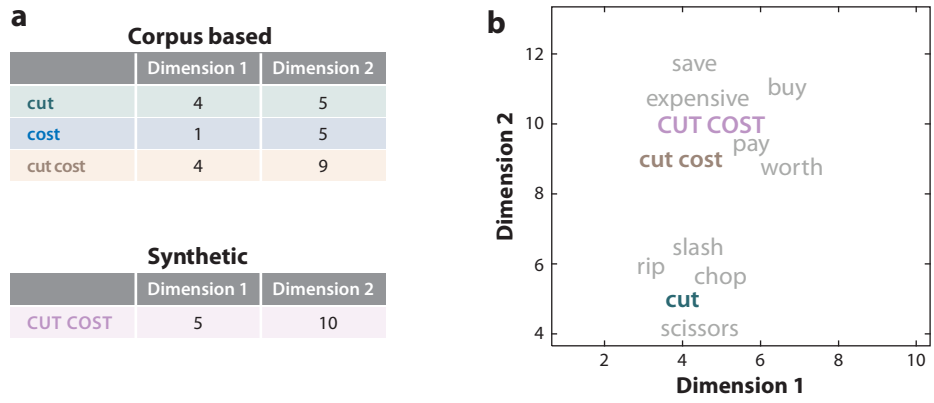

Figure 3

Compositional distributional semantics: illustration with vector addition. (*a*) The synthetic vector CUT COST is built by component-wise addition of the vectors for *cut* and *cost*. (*b*) The argument *cost* pulls the vector for *cut* toward its abstract use (see nearest neighbors in gray). The corpus-based vector for *cut cost* can be used to check the quality of its synthetic counterpart.

as a result of the composition (Baroni et al. 2014a). **Figure 3** provides an illustration using the simplest composition method: adding the word vectors. The synthetic vector CUT COST created via this composition method has a value of $4 + 1 = 5$ for the first dimension because the values of *cut* and *cost* for this dimension are 4 and 1, respectively.

To appreciate how this method may account for semantic effects, let us assume that dimension 1 is associated with concrete notions and dimension 2 with abstract notions (of course, this is a simplification; recall that properties such as concreteness are captured in a distributed fashion). The verb *cut* has a concrete sense, as in *cut paper*, and a more abstract sense akin to *save*, as in *cut costs*, so it has high values for both dimensions. By contrast, *cost* is an abstract notion, so it has low values for dimension 1 and high values for dimension 2. When composing the two, the abstract dimension is highlighted, pulling the vector toward regions in the semantic space related to its abstract sense. **Figure 3*b*** illustrates: While the vector values are fictitious, the neighbors are a selection of the 20 nearest neighbors of *cut* and CUT COST in a real semantic space (Mandera et al. 2017).[2] As the nearest neighbors show, the representation of *cut* is dominated by the physical sense, but its composition with *cost* shifts it toward the abstract sense. The mechanism by which matching semantic dimensions reinforce one another, and mismatched dimensions remain less active, is reminiscent of the mechanisms presented by Pustejovsky (1995), discussed above, for *bake a potato* versus *bake a cake*. The main difference is that distributional representations are not explicitly structured like those in the Generative Lexicon, although that does not mean that they lack structure but, rather, that the structure is implicitly defined in the space.

A substantial body of research has shown that composition methods in distributional semantics largely account for polysemy effects in semantic composition. Baroni & Zamparelli (2010) introduce an evaluation method to assess how good composed representations are. They compare the synthetic vector for a phrase, such as CUT COST in **Figure 3**, to a phrase vector, *cut cost*, that is extracted directly from the corpus with standard distributional methods. The closer the synthetic vectors are to the corpus-based ones, the better the composition method is. In a study by Boleda et al. (2013), the best composition method obtains an average cosine similarity of 0.6 between synthetic and corpus-based vectors for adjective–noun phrases; for comparison, phrases have an

[2] A user-friendly interface to this semantic space can be found at **http://meshugga.ugent.be/snaut-english**.

average cosine similarity to their head nouns of 0.4. Another common method is to compare model results with human intuitions about the semantics of phrases. Mitchell & Lapata (2010) introduced this method for phrase similarity (which, in turn, was inspired by methods to evaluate word similarity), with participant data such as *reduce amount–cut cost* being very similar, *encourage child–leave company* being very dissimilar, and *present problem–face difficulty* obtaining medium scores. The best composition methods yield Spearman correlation scores with participant data around 0.4 (minimum is 0, maximum 1) for adjective–noun, noun–noun, and verb–noun phrases; for comparison, correlation scores between different participants are around 0.5. Other research has experimented with ditransitive constructions (Grefenstette & Sadrzadeh 2011), with triples such as *medication achieve result–drug produce effect*, or even full sentences (Bentivogli et al. 2016), but going beyond short phrases proves difficult. There is limited research on function words because, as mentioned above, these are traditionally hard to model with distributional semantics. An exception is a study by Bernardi et al. (2013), who seek to identify paraphrasing relationships between determiner phrases (e.g., *several wives*) and words that lexically involve some form of quantification (e.g., *polygamy*). They obtain reasonable but not optimal results.

A particularly exciting application of compositional distributional methods is that of Vecchi et al. (2017), who showed that distributional models can distinguish between semantically acceptable and unacceptable adjective–noun phrases. Crucially, their data involve phrases that are unattested in a very large corpus; some phrases are unattested because they are semantically anomalous (e.g., *angry lamp*, *legislative onion*), and some because of the generative capacity of language, with its explosion of combinatory properties, together with the properties of the world, which make some combinations of adjectives and nouns unlikely even if they are perfectly acceptable (e.g., *warm garlic*, *sophisticated senator*). The fact that distributional models are able to predict which combinations are acceptable for human participants, and which are not, suggests that they are able to truly generalize.

Researchers in this area have investigated approaches to composition much more sophisticated than vector addition. I cannot do justice to this research for reasons of space (but see Baroni 2013, Erk 2012). Much of it is inspired by formal semantics (Baroni et al. 2014a, Beltagy et al. 2013, Coecke et al. 2011, Erk 2016, Garrette et al. 2011, Herbelot & Vecchi 2015, Lewis & Steedman 2013); Boleda & Herbelot (2016) review research at the intersection between formal and distributional semantics. However, a robust result that has emerged from all this literature is that vector addition is surprisingly good, often outperforming more sophisticated methods. This finding suggests that the greatest power of distributional semantics lies in the lexical representations themselves.

Relatedly, this research has also shown that, while distributional semantics can model composition of content words in short phrases, scaling up to larger constituents and accounting for function words remain challenging. Recall that distributional semantics provides abstractions over all occurrences of an expression. Compositionally built phrases remain generic rather than grounded to a specific context. Therefore, distributional approaches can account for the fact that, in general, *red box* will be used for boxes that are red in color, but they cannot really account for highly context-dependent interpretations, such as *red box* referring to a brown box containing red objects (McNally & Boleda 2017). This is because distributional semantics does not come equipped with a mechanism to integrate word meaning in a given linguistic and extralinguistic context, or to represent that context in the first place (Aina et al. 2019). Note that functional elements such as tense or determiners need a context to be interpreted, so it makes sense that they are challenging for distributional semantics.

Accordingly, Westera & Boleda (2019) defend the view that distributional semantics accounts for expression meaning (more concretely, how an expression is typically used by speakers), but not

for speaker meaning (how a speaker uses an expression, in terms of communicative intentions, in a given context; see the *red box* example, above). Newer-generation neural networks are contributing to expanding these limits, as they natively incorporate mechanisms to compose new words with a representation of the context (Mikolov et al. 2010). However, the extent to which they can account for speaker meaning, and contextual semantic effects more generally, remains to be determined; results obtained by Aina et al. (2019) suggest that some of these models still overwhelmingly rely on lexical information.

Finally, another area where distributional semantics shows potential is the phenomenon of semantic opacity and semiopacity, which is the opposite of compositionality. I discuss research on the compositionality of noun compounds in Section 4.2, below.

3.2. Different Representations, Polysemy via Word Senses

Other research aims to build sense-specific distributional representations, where typically each word sense is assigned a different vector (for a recent survey, see Camacho-Collados & Pilehvar 2018). The key insight here is that, because distributional semantics is based on context of use and uses of a word in a given sense will be more similar to one another than to uses of the same word in a different sense, we can detect word senses by checking the similarity of the contexts. Pioneering research by Schütze (1998) did so by representing single instances of word use, with one vector for each sentence in which the word occurs. Then, word senses were automatically identified as coherent regions in that space. Schütze (1998) started a tradition, within distributional semantics, of research on word sense induction and sense-specific word representations (McCarthy et al. 2004, Reisinger & Mooney 2010). By contrast, Erk and colleagues aimed to provide a representation of the specific meaning a word takes in a given context, entirely bypassing word senses (Erk & Padó 2008, 2010; Erk et al. 2013). This research is related to compositional distributional semantics, with the difference that it provides a use-specific word vector representation instead of going directly to the representation of the larger constituent.

Two crucial problems in sense-based approaches to polysemy are (*a*) determining when two senses are different enough to warrant the addition of an item to the vocabulary and (*b*) deciding how to represent the information that is common to different senses (Kilgarriff 1997). Distributional semantics does not improve things with respect to the first issue, but it does alleviate the second. Two sense-specific vectors can be similar in some dimensions (e.g., for *cut*, those related to reducing or splitting) and different in others (e.g., the abstract/concrete axis of *cut*), in a graded fashion. In the same way that distributional semantics can capture similarities and differences between words, it can capture similarities and differences between word senses.

3.3. Discussion

Polysemy is a pervasive phenomenon that is difficult to model in a discrete, symbolic system (Kilgarriff 1997). Distributional semantics provides an attractive framework, complementary to traditional ones (Heylen et al. 2015). Multidimensionality allows it to capture both the common core in different uses of a word and the differential factors, as some dimensions of meaning can specialize in the former and some in the latter. Gradedness allows it to capture the degree of the semantic shift in different contexts of use, be it in the composition route or the word sense route. Moreover, the fact that distributional semantics provides data-induced representations for a large number of words makes it possible to use it to make predictions and test specific hypotheses driven by linguistic theory.

As an example, Boleda et al. (2013) test the hypothesis, stemming from formal semantics, that modification by a certain class of adjectives is more difficult to model than other classes. The specific prediction is that synthetic phrases with these adjectives (e.g., ALLEGED KILLER) will be further away from their corpus-based vectors than synthetic phrases with other adjectives (e.g., SEVERE PAIN). Their results are negative, and they instead observe the influence of another factor in the results: If an adjective denotes a very typical property of a noun, as in *severe* for *pain*, then it is easy to model; if it is less typical, as in *severe* for *budget*, then it is more difficult. In many of the difficult cases, such as *likely base*, it is not even clear how the two words compose; out of context, it is not easy to come up with possible interpretations for this phrase. This led the authors to further explore the context dependence of modification, resulting in a theoretical proposal (McNally & Boleda 2017). These authors propose that composition exploits two aspects of meaning: on the one hand, the conceptual aspect, with regularities in how words match (e.g., *box* denotes a physical object, and *red* is typically used to specify colors of physical objects) and, on the other hand, the referential aspect, specifically the information about the referent of the phrase (e.g., *red box* can be used in a context that requires distinguishing a brown box containing red objects from another, identical-looking brown box containing blue objects). Distributional semantics can model conceptual but not referential effects, for the same reason that it cannot model contextual effects and speaker meaning more generally (see Section 3.1). McNally & Boleda (2017) take distributional semantic data themselves as an object of empirical inquiry; they ask what makes certain phrases difficult for compositional distributional models, and the answer proves theoretically worthy. This research thus represents an example of fruitful collaboration between computational and theoretical approaches to language.

4. GRAMMAR–SEMANTICS INTERFACE

There is ample evidence that content-related aspects of language interact with formal aspects, as is salient in, for instance, argument structure and the expression of arguments in syntax (Grimshaw 1990, Levin 1993; see Section 4.1), as well as in derivational morphology (Lieber 2004; see Section 4.2).

4.1. Syntax–Semantics Interface

Beth Levin's seminal research on the syntax–semantics interface was based on the observation that "the behavior of a verb, particularly with respect to the expression of its arguments, is to a large extent determined by its meaning" (Levin 1993, p. 1). She defines semantic verb classes on the basis of several syntactic properties. This is a particular case of the Distributional Hypothesis; thus, it is natural to turn it around and use distributional cues to infer semantic classes—as Levin herself does manually.

Levin's work had a significant impact on computational linguistics, inspiring research on the automatic acquisition of semantic classes from distributional evidence (Boleda et al. 2012, Dorr & Jones 1996, Korhonen et al. 2003, Lapata & Brew 2004, McCarthy 2000, Merlo & Stevenson 2001, Schulte im Walde 2006). For instance, Merlo & Stevenson (2001) used manually defined linguistic features, with data extracted from corpora, to classify English verbs into three optionally transitive classes: unergative, unaccusative, and object-drop. They achieved around 70% accuracy. Other research has targeted a finer-grained classification, with Levin-style semantic classes, as Schulte im Walde (2006) did for German. This early work used distributional evidence, but not distributional semantics strictu sensu. Baroni & Lenci (2010) replicated Merlo & Stevenson's (2001) experiment by using a proper distributional model, obtaining comparable accuracy.

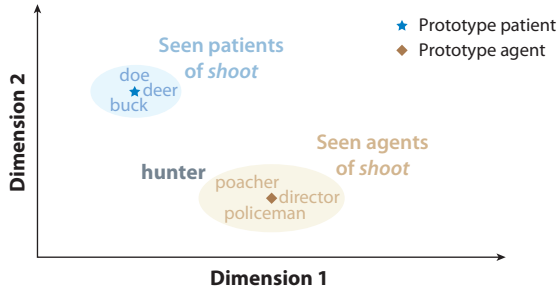

Figure 4

The approach to selectional preferences by Erk et al. (2010): The plausibility of verb–argument combinations is measured in terms of the similarity between a candidate argument and a prototype representation of its arguments. The prototype is a weighted average of the arguments observed to occur with a given verb in a corpus. In this example, *hunter* is a plausible agent for *shoot*, rather than a patient, because its vector is closer to the prototype agent vector than the prototype patient vector of the verb. The prototype agent vector, in turn, has been computed as a weighted average of the observed agents *poacher*, *director*, and *policeman*. Figure adapted with permission from Erk et al. (2010, p. 731, figure 1) under a Creative Commons Attribution (CC BY) 3.0/4.0 License.

Erk et al. (2010) initiated a line of research that used distributional methods to model selectional restrictions, or the thematic fit between an argument and a predicate (usually a verb). They capitalized on the fact that distributional models capture gradedness in linguistic phenomena, since selectional restrictions are graded: *cake* is a better object for *eat* than *chalk*, which, in turn, is better than *sympathy*. Again, this gradedness is not easy to capture in symbolic models with discrete features like [±EDIBLE]. Erk et al. computed the plausibility of each verb–argument combination as the similarity between a candidate argument and a (weighted) average of the arguments observed with a verb. For instance, when deciding whether *hunter* is a plausible agent for the verb *shoot*, they computed its similarity to an average of the vectors for *poacher*, *director*, *policeman*, and so forth (**Figure 4**). This average vector can be regarded as a prototype for the argument of the verb.

Erk et al. compared the scores of the model with human ratings (where participants were asked to rate the plausibility that, e.g., *hunter* is an agent of *shoot*). Their model achieved Spearman correlation scores of 0.33 and 0.47 ($p < 0.001$) with the human ratings in two different data sets for English involving agent and patient roles. Erk et al.'s idea of working with argument prototypes has been further refined and developed in subsequent models (Baroni & Lenci 2010, Greenberg et al. 2015, Lenci 2011, Santus et al. 2017), with improved empirical results and a broader coverage of phenomena.

4.2. Morphology–Semantics Interface

Derivational morphology is at the interface between grammar and semantics (Lieber 2004). Stem and affix need to match in both morphosyntactic and semantic features. For instance, the suffix *-er* applies to verbs, as in *carve* → *carver*, but only those that have certain kinds of arguments. The effects of derivational processes are also both grammatical (*-er* produces nouns) and semantic (these nouns, e.g., *carver*, have some agentive connotation). Derivational processes are semiregular; they are largely compositional, but not always (mainly due to lexicalization processes), and they present subregularities (e.g., *carver* and *driver* denote agents, but *broiler* and *cutter* denote instruments). Moreover, both stem and affix semantics exhibit the properties typical of word semantics,

such as polysemy and gradedness (Marelli & Baroni 2015); compare with the polysemy of *-er* between agent and instrument. Thus, accounting for morphological derivation requires fine-grained lexical semantic representations for both stem and affix as well as mechanisms to combine them, in a clear analogy to phrase composition (see Section 3.1).

In recent years, researchers have explored methods to produce distributional representations for morphologically complex words from the representations of their parts (e.g., Cotterell & Schütze 2018, Lapesa et al. 2018, Lazaridou et al. 2013, Marelli & Baroni 2015, Padó et al. 2016); most of this work has adapted compositional methods initially developed for word composition. The motivation of this research is twofold. From a theoretical point of view, distributional semantics offers new tools to investigate derivational morphology, in particular, its rich, data-driven semantic representations. From a practical perspective, such methods address "the data problem" of distributional semantics (Padó et al. 2016, p. 1285): In general, morphologically complex words are less frequent than morphologically simple words, so distributional representations for morphologically complex words can be expected to be of a comparatively lower quality. Moreover, because morphology is productive, new words are continuously created, and in these cases data are simply unavailable, so it is imperative to rely on methods to build synthetic word vectors for morphologically complex words.

Researchers have experimented with simple composition methods and more complex ones, often based on machine learning. Again, the simplest method is addition, which here implies summing the vectors for the stem and the affix, as in CARVER = *carve* + *-er*. However, affixes are not observed as units in corpora. A common method of obtaining affix representations is to average derived words (Padó et al. 2016)—for instance, averaging the vectors for *carver*, *drinker*, *driver*, and so forth to obtain a representation for *-er*.

Table 2 lists phenomena captured by the distributional model presented by Marelli & Baroni (2015), illustrated through nearest neighbors (see their original paper for quantitative evaluation; also note that their composition method is more sophisticated than addition, but the kinds of effects modeled are similar for different composition functions). The first row of the table shows that the distributional method captures the agent/instrument polysemy of the affix and is able to produce different results depending on the stem: The synthetic vector for CARVER is near agents for professions (*potter*, *goldsmith*), whereas BROILER is in the region of cooking instruments (*oven*, *stove*). The second row shows that the relevant sense of the stem is captured even in cases where it is not the predominant one: In the vector for the word *column*, the senses related to architecture and mathematics dominate (see the nearest neighbors), but *-ist* correctly focuses on the sense related to

Table 2 Derivational phenomena captured with compositional distributional semantic methods[a]

Phenomenon	Word[b]	Nearest neighbors (selection)[c]
Affix polysemy	CARVER	potter, engraver, goldsmith
	BROILER	oven, stove, to cook, kebab, done
Sense selection	column	arch, pillar, bracket, numeric
	COLUMNIST	publicist, journalist, correspondent
Differential effect of the affix	INDUSTRIAL	environmental, land-use, agriculture
	INDUSTRIOUS	frugal, studious, hardworking

[a]Examples are from Marelli & Baroni (2015).
[b]The synthetic word representations in small capital letters are produced by derivation operations with distributional semantics; the words in regular font correspond to corpus-based word vectors.
[c]Marelli & Baroni (2015) provide a selection of the 20 nearest neighbors.

journalism when producing COLUMNIST. Because *-ist* often produces professions, its distributional representation is able to select the dimensions of *column* that match one of the meaning types produced by the morpheme. Finally, the examples in the third row show that different affixes produce different meanings when applied to the same stem. For instance, -AL and -OUS have quite different consequences on the same base form.

Distributional semantics has clear potential to capture linguistic phenomena related to derivation; the extent to which it is able to do so is still under investigation, since distributional methods exhibit a wide variance in performance across individual words and across derivational patterns (Padó et al. 2016). The intervening factors are still not fully understood, but it seems clear that some are methodological and some are linguistic. As for the former, if a word is very frequent, it will have probably undergone lexicalization; if it is very infrequent, then its corpus-based representation will be of low quality. In both cases, the word will not be a good candidate to participate in the creation of the affix representation, or as a comparison point to evaluate distributional methods. Thus, it is not surprising that overall scores are good but not optimal. For instance, Lazaridou et al. (2013), in a study on English, showed that derived forms have a mean cosine similarity of 0.47 with their base forms (e.g., *carver* compared with *carve*). The best compositional measure provides a mean similarity of 0.56 between synthetic and corpus-based vectors—significantly higher, but not a big jump. However, they also provide evidence that, in cases where the quality of the corpus-based word representations is low, the compositional representation is substantially better, suggesting that distributional methods can provide useful semantic representations for derived words in a productive way and can alleviate the data problem explained above. For instance, the nearest neighbors of *rename* in their distributional space are *defunct*, *officially*, and *merge*, whereas those for the synthetic vector RENAME are *name*, *later*, and *namesake*.

As for linguistic factors, Padó et al. (2016), in a large-scale study of derivation in German, find that the derivational pattern is the best predictor of model performance (i.e., some derivational processes are intrinsically harder to model than others) and argue that derivations that create new argument structure tend to be harder for distributional models. For instance, the agentive/instrumental nominalization with suffix *-er* (*fahren* → *Fahrer*, English *drive–driver*), where the external argument is incorporated into the word, is difficult to capture, whereas deverbal nominalizations that preserve argument structure are comparatively easy (e.g., with suffix *-ung*; *umleiten* → *Umleitung*, English *redirect–redirection*).

Research in derivational morphology also shows that vector addition works surprisingly well, as is the case with composition (see Section 3.1). This finding, again, suggests that the distributional representations themselves do most of the job and are more important than the specific method used. The results obtained by Cotterell & Schütze (2018) underscore this interpretation. These authors propose a probabilistic model that integrates the automatic decomposition of words into morphemes (*carver* → [*carve*] [*er*]) with the synthesis of their word meaning, jointly learning the structural and semantic properties of derivation. They test different derivation models and different word representations on English and German data, with representations having by far the most influence on the results.

The robustness of addition has also emerged in the study of semantic opacity and semiopacity, which typically aims to predict the degree of compositionality in compound nouns and multiword expressions. In a representative study, Reddy et al. (2011) sought to reproduce human ratings on the degree of compositionality of 90 English compound nouns (*climate change*, *graduate student*, and *speed limit* obtained maximum compositionality scores; *silver bullet*, *ivory tower*, and *gravy train*, minimum). Adding the two component vectors (with a higher weight of the modifier; see Reddy et al. 2011 for details) yielded a Spearman correlation score of 0.71 with human data. Other research uses different methods; for instance, Springorum et al. (2013) do not use compositional

methods but rather explore how the modifier and the head contribute to compositionality ratings for German data. In contrast to their prediction, the modifier is a much better predictor of compositionality than the head.

Again, most research is directed at showing that distributional semantics can model derivational morphology, rather than tackling more specific linguistic questions. An exception is a study by Lapesa et al. (2017), an interdisciplinary collaboration between theoretical and computational linguists that tests hypotheses about the effect of derivation on emotional valence (the positive or negative evaluation of the referent of a word) on German data. For instance, one of the study's predictions is that diminutives shift words toward positive valence (consider *Hund* → *Hündchen*, English *dog–doggie*). This study provides overall support for the hypotheses, but in a nuanced form. The most interesting result is a hitherto-unobserved interaction of many valence effects with concreteness: The diminutive makes nouns positive if they denote concrete objects, whereas it tends to make abstract nouns negative (compare the case of *dog* with *Idee* → *Ideechen*, English *idea–small idea*), and verbal prefixation with *über-* (*over-*) tends to make concrete verbs, but not abstract verbs, negative (*fahren* → *überfahren*, English *drive–run over* versus *nehmen* → *übernehmen*, English *take–take over*). This research, again, demonstrates the potential of distributional semantics to uncover linguistically relevant factors.

Although most work is on derivational morphology, some research has tackled inflection, too. A very influential study, with a model first proposed by Rumelhart & Abrahamson (1973), is that by Mikolov et al. (2013b), who showed that several morphological and semantic relations are organized according to simple additive relationships in distributional space. For instance, *good–better+rough* creates a synthetic vector that is very near *rougher*. The idea is that if one subtracts an inflected word from its stem, one obtains a representation of the affix (here, the comparative), which can then be applied to a new stem (here, *rough*) by addition. Mikolov et al. tested eight patterns involving nominal, adjectival, and verbal inflection, obtaining an average accuracy of 40% on the task of predicting the missing element in the tuple. This value may not seem very impressive, but it is if we consider that the model is required to find the exact right answer in a vocabulary of 82,000 words—that is, in the case above, the answer is counted as correct only if the nearest neighbor of the synthetic vector ROUGHER is the word vector of *rougher*.

4.3. Discussion

The literature reviewed in this section has three main assets to offer to theoretical approaches to the grammar–semantics interface. The first is a wealth of data, created as part of the research in order to develop and evaluate distributional methods. For example, participant ratings on the compositionality of compounds (Reddy et al. 2011) can be used when selecting material for experimental research. Other examples are ratings of typicality and semantic relatedness (Lazaridou et al. 2013, Springorum et al. 2013) and information about derived words, such as derivational pattern and degree of polysemy (Padó et al. 2016). This kind of contribution is common to other quantitative and computational work (Baayen et al. 1993).

The second is tools to create and explore data via distributional methods. For instance, the similarity between a derived form and a combination of its components in distributional space can be used as a proxy for its degree of compositionality, which is useful to explore processes of derivation and lexicalization. Other linguistic features can be simulated with distributional measures. For instance, Padó et al. (2016) measure how semantically typical a base form is for a given morphological pattern by comparing it with the average of all the bases in the pattern (e.g., for *-er*, the word vector for *carve* compared with the average of the vectors for *carve*, *drink*, *drive*, *broil*, *cut*, and so on).

The third is the potential to uncover new empirical facts that are of potential theoretical significance. Examples include the suggestion by Padó et al. (2016) that derivation processes that affect argument structure are more challenging to model computationally, as well as the relevance of the concreteness/abstractness axis in the study by Lapesa et al. (2017).

5. CONCLUSION AND OUTLOOK

This review summarizes robust results in distributional semantics that can be directly imported into research in theoretical linguistics, as well as challenges and open issues. Among these results are that (*a*) distributional semantics is particularly useful in areas where the connection among use, meaning, and grammar is relevant, such as the areas reviewed in this article; (*b*) geometric relationships in distributional models correspond to semantic relationships in language; (*c*) gradedness in distributional representations correlates with gradedness in semantic phenomena (e.g., the degree of semantic change); (*d*) averaging the distributional representations of classes of words yields useful abstractions of the relevant classes (e.g., of arguments accepted by specific predicates); and (*e*) simple combinations of distributional representations produce quite accurate predictions regarding the semantics of phrases and derived words. I have argued that the multidimensional, graded, and data-driven nature of representations in distributional semantics are key aspects that contribute to these results.

There are at least four ways for distributional semantic research to contribute to linguistic theories. The first is exploratory. Distributional data such as similarity scores and nearest neighbors can be used to explore data on a large scale. The second is as a tool to identify instances of specific linguistic phenomena. For instance, changes in distributional representations of words across time can be used to systematically harvest potential instances of semantic change in diachronic data (Section 2). The third is as a test bed for linguistic hypotheses, by testing predictions in distributional terms. The fourth, and hardest, is the actual discovery of linguistic phenomena or theoretically relevant trends in data. For actual discoveries to be made, collaboration between computational and theoretical linguists is needed.

Distributional methods face a number of challenges. Similar to other data-driven methods, distributional models mirror the data they are fed. This is good, because the models provide radically empirical representations, and it is also dangerous, because representations are subject to biases in the underlying data (Caliskan et al. 2017). A related challenge is that distributional methods need large quantities of data to learn reasonable representations. A rule of thumb is to have at least 20–50 instances of each expression to represent; many languages, domains, or time periods simply lack these data. There is active research on faster learning, as this is a problem for many other areas, but no working solution at present. A final, crucial issue is the lack of adequate researcher training, which prevents wider use of distributional semantics in linguistics and of quantitative and computational methods more generally. Strengthening training of linguistics students in such methods will be paramount to allow the field to adequately exploit the vast quantity of linguistic data that has become available in the last few decades.

In this review, to maximize readability, I have focused on simple methods such as vector similarity, nearest neighbors, vector addition, and vector averaging. While these are the basic methods in the field, a glaring omission is methods based on machine learning techniques, which are also commonly used to extract information from semantic spaces and operate with distributional representations. I refer the reader to the references provided above for more information.

For reasons of scope, I have also omitted research on neural networks that is not specifically targeted at building semantic spaces. Neural networks are a type of machine learning algorithm,

recently revamped as deep learning (LeCun et al. 2015), that induce representations of the data they are fed in the process of learning to perform a task. For instance, they learn word representations as they learn to translate from English to French, given large amounts of bilingual text. They proceed by trial and error, attempting to translate a sentence, measuring the degree of error, and feeding back to the representations such that they become more helpful for the task. Linguistic tasks that are general enough, such as machine translation or word prediction, result in general-purpose representations of language. Most deep learning systems for language include a module that is akin to a distributional lexicon, and everything I have reported in this review applies to such modules. However, crucially, these systems also have other modules that represent linguistic context, as well as mechanisms to combine this context with word representations. This is a big step with respect to classic distributional models, and deep learning is being adopted in the community at top speed. To illustrate, examples 3 and 4 (Peters et al. 2018, p. 2233) show the nearest neighbors of two sentences containing the polysemous word *play*, where the representations for the sentences are vectors produced by the complex compositional function implemented in the neural network. The nearest neighboring sentences illustrate that the model has captured not only the relevant sense of the word but also more nuanced aspects of the meanings of the sentences (commenting on good plays in example 3, referring to signing for plays rather than to the acting itself in example 4):

(3) Sentence: Chico Ruiz made a spectacular play on Alusik's grounder […].
Nearest neighbor: Kieffer […] was commended for his ability to hit in the clutch, as well as his all-round excellent play.

(4) Sentence: Olivia de Havilland signed to do a Broadway play for Garson […].
Nearest neighbor: […] they were actors who had been handed fat roles in a successful play […].

Given the success (and complexity) of these models, there is booming interest in the computational linguistic community in understanding what aspects of language they capture, and how (Alishahi et al. 2019). Recently, Pater (2019) argued for the integration of neural network models in linguistic research. I could not agree more.

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 am grateful to Louise McNally, Josep Maria Fontana, Alessandro Lenci, and Marco Baroni for discussions about the role of distributional semantics in linguistic theory; to the AMORE team (Laura Aina, Kristina Gulordava, Carina Silberer, Ionut Sorodoc, and Matthijs Westera) for shaping my current thinking about this topic; and to Dan Jurafsky for a helpful review of a previous draft. This project has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (grant 715154) and from the Spanish Ramón y Cajal program (grant RYC-2015-18907). This review reflects the author's view only, and the European Union is not responsible for any use that may be made of the information it contains.

LITERATURE CITED

Aina L, Gulordava K, Boleda G. 2019. Putting words in context: LSTM language models and lexical ambiguity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3233–48. Stroudsburg, PA: Assoc. Comput. Linguist.

Alishahi A, Chrupała G, Linzen T. 2019. Analyzing and interpreting neural networks for NLP: a report on the first BlackboxNLP Workshop. arXiv:1904.04063 [cs.CL]

Arora S, Li Y, Liang Y, Ma T, Risteski A. 2018. Linear algebraic structure of word senses, with applications to polysemy. *Trans. Assoc. Comput. Linguist.* 6:483–95

Baayen RH, Piepenbrock R, Gulikers L. 1993. *The CELEX lexical database*. CD-ROM, Linguist. Data Consort., Philadelphia

Baroni M. 2013. Composition in distributional semantics. *Lang. Linguist. Compass* 7:511–22

Baroni M. 2016a. *Composes: an executive summary*. Talk presented at the Composes Workshop, Florence, Italy, Aug. 14

Baroni M. 2016b. Grounding distributional semantics in the visual world. *Linguist. Lang. Compass* 10:3–13

Baroni M, Bernardi R, Zamparelli R. 2014a. Frege in space: a program for 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, 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* (*EMNLP 2010*), pp. 1183–93. Stroudsburg, PA: Assoc. Comput. Linguist.

Beltagy I, Chau C, Boleda G, Garrette D, Erk K, Mooney R. 2013. Montague meets Markov: deep semantics with probabilistic logical form. In *2nd Joint Conference on Lexical and Computational Semantics* (**SEM*), Vol. 1: *Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pp. 11–21. Stroudsburg, PA: Assoc. Comput. Linguist.

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

Bernardi R, Dinu G, Marelli M, Baroni M. 2013. A relatedness benchmark to test the role of determiners in compositional distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pp. 53–57. Stroudsburg, PA: Assoc. Comput. Linguist.

Bojanowski P, Grave E, Joulin A, Mikolov T. 2017. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* 5:135–46

Boleda G, Baroni M, Pham TN, McNally L. 2013. Intensionality was only alleged: on adjective–noun composition in distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics* (*IWCS 2013*), pp. 35–46. Stroudsburg, PA: Assoc. Comput. Linguist.

Boleda G, Erk K. 2015. Distributional semantic features as semantic primitives—or not. In *Papers from the AAAI Spring Symposium. Knowledge Representation and Reasoning: Integrating Symbolic and Neural Approaches*, 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

Boleda G, Schulte im Walde S, Badia T. 2012. Modeling regular polysemy: a study on the semantic classification of Catalan adjectives. *Comput. Linguist.* 38:575–616

Bréal M. 1897. *Essai de sémantique*. Paris: Hachette

Caliskan A, Bryson JJ, Narayanan A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356:183–86

Camacho-Collados J, Pilehvar MT. 2018. From word to sense embeddings: a survey on vector representations of meaning. *J. Artif. Intell. Res.* 63:743–88

Clark S. 2015. Vector space models of lexical meaning. In *The Handbook of Contemporary Semantic Theory*, ed. S Lappin, C Fox, pp. 493–522. New York: Wiley

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

Cotterell R, Schütze H. 2018. Joint semantic synthesis and morphological analysis of the derived word. *Trans. Assoc. Comput. Linguist.* 6:33–48

Cruse DA. 1986. *Lexical Semantics*. Cambridge, UK: Cambridge Univ. Press

Davies M. 2010. *The Corpus of Historical American English* (*COHA*). **https://www.english-corpora.org/coha/**

Del Tredici M, Fernández R, Boleda G. 2019. Short-term meaning shift: a distributional exploration. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (*NAACL-HLT 2019*), pp. 2069–75. Stroudsburg, PA: Assoc. Comput. Linguist.

Deo A. 2015. Diachronic semantics. *Annu. Rev. Linguist.* 1:179–97

Dorr BJ, Jones D. 1996. Role of word sense disambiguation in lexical acquisition: predicting semantics from syntactic cues. In *Proceedings of the 16th Conference on Computational Linguistics* (*COLING96*), Vol. 1, pp. 322–27. Stroudsburg, PA: Assoc. Comput. Linguist.

Dubossarsky H, Weinshall D, Grossman E. 2017. Outta control: laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (*EMNLP 2017*), pp. 1136–45. Stroudsburg, PA: Assoc. Comput. Linguist.

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, McCarthy D, Gaylord N. 2013. Measuring word meaning in context. *Comput. Linguist.* 39:511–54

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* (*EMNLP 2008*), pp. 897–906. Stroudsburg, PA: Assoc. Comput. Linguist.

Erk K, Padó S. 2010. Exemplar-based models for word meaning in context. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 92–97. 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

Fillmore CJ. 2006. Frame Semantics. In *Cognitive Linguistics: Basic Readings*, ed. D Geeraerts, pp. 373–400. Berlin: Mouton de Gruyter

Garrette D, Erk K, Mooney R. 2011. Integrating logical representations with probabilistic information using Markov logic. In *Proceedings of the 9th International Conference on Computational Semantics* (*IWCS 2011*), pp. 105–14. Stroudsburg, PA: Assoc. Comput. Linguist.

Greenberg C, Sayeed A, Demberg V. 2015. Improving unsupervised vector-space thematic fit evaluation via role-filler prototype clustering. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (*NAACL-HLT 2015*), pp. 21–31. Stroudsburg, PA: Assoc. Comput. Linguist.

Grefenstette E, Sadrzadeh M. 2011. Experimenting with transitive verbs in a DisCoCat. In *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics*, pp. 62–66. Stroudsburg, PA: Assoc. Comput. Linguist.

Grimshaw J. 1990. *Argument Structure*. Cambridge, MA: MIT Press

Gulordava K, Baroni M. 2011. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics*, pp. 67–71. Stroudsburg, PA: Assoc. Comput. Linguist.

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. 1489–501. Stroudsburg, PA: Assoc. Comput. Linguist.

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

Herbelot A, Vecchi EM. 2015. Building a shared world: mapping distributional to model-theoretic semantic spaces. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (*EMNLP 2015*), pp. 22–32. Stroudsburg, PA: Assoc. Comput. Linguist.

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

Hock HH. 1991. *Principles of Historical Linguistics*. Berlin: Walter de Gruyter

Kilgarriff A. 1997. I don't believe in word senses. *Comput. Humanit.* 31:91–113

Kim Y, Chiu YI, Hanaki K, Hegde D, Petrov S. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pp. 61–65. Stroudsburg, PA: Assoc. Comput. Linguist.

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

Korhonen A, Krymolowski Y, Marx Z. 2003. Clustering polysemic subcategorization frame distributions semantically. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, Vol. 1, pp. 64–71. Stroudsburg, PA: Assoc. Comput. Linguist.

Kutuzov A, Øvrelid L, Szymanski T, Velldal E. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics* (*COLING 2018*), pp. 1384–97. Stroudsburg, PA: Assoc. Comput. Linguist.

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

Lapata M, Brew C. 2004. Verb class disambiguation using informative priors. *Comput. Linguist.* 30:45–73

Lapesa G, Kawaletz L, Plag I, Andreou M, Kisselew M, Padó S. 2018. Disambiguation of newly derived nominalizations in context: a distributional semantics approach. *Word Struct.* 11:277–312

Lapesa G, Padó S, Pross T, Roßdeutscher A. 2017. Are *doggies* cuter than *dogs*? Emotional valence and concreteness in German derivational morphology. In *Proceedings of the 12th International Conference on Computational Semantics* (*IWCS 2017*), pp. 1–7. Stroudsburg, PA: Assoc. Comput. Linguist.

Lazaridou A, Marelli M, Zamparelli R, Baroni M. 2013. Compositional-ly derived representations of morphologically complex words in distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pp. 1517–26. Stroudsburg, PA: Assoc. Comput. Linguist.

LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. *Nature* 521:436–44

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. 2018. Distributional models of word meaning. *Annu. Rev. Linguist.* 4:151–71

Levin B. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago/London: Univ. Chicago Press

Lewis M, Steedman M. 2013. Combined distributional and logical semantics. *Trans. Assoc. Comput. Linguist.* 1:179–92

Lieber R. 2004. *Morphology and Lexical Semantics*. Cambridge, UK: Cambridge Univ. Press

Lyons J. 1977. *Semantics*. Cambridge, UK: Cambridge Univ. Press

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

Marelli M, Baroni M. 2015. Affixation in semantic space: modeling morpheme meanings with compositional distributional semantics. *Psychol. Rev.* 122:485–515

McCarthy D. 2000. Using semantic preferences to identify verbal participation in role switching alternations. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference* (*NAACL 2000*), pp. 256–63. Stroudsburg, PA: Assoc. Comput. Linguist.

McCarthy D, Koeling R, Weeds J, Carroll J. 2004. Finding predominant word senses in untagged text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pp. 279–86. Stroudsburg, PA: Assoc. Comput. Linguist.

McNally L, Boleda G. 2017. Conceptual versus referential affordance in concept composition. In *Compositionality and Concepts in Linguistics and Psychology*, ed. JA Hampton, Y Winter, pp. 245–67. Berlin: Springer

Merlo P, Stevenson S. 2001. Automatic verb classification based on statistical distributions of argument structure. *Comput. Linguist.* 27:373–408

Michel JB, Shen YK, Aiden AP, Veres A, Gray MK, et al. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331:176–82

Mikolov T, Karafiat M, Burget L, Cernocky J, Khudanpur S. 2010. Recurrent neural network based language model. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association* (*INTERSPEECH 2010*), pp. 1045–48. Baixas, Fr.: Int. Speech Commun. Assoc.

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. 2013a. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems* (*NIPS13*), Vol. 2, pp. 3111–19. Red Hook, NY: Curran

Mikolov T, Yih W, Zweig G. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (*NAACL 2013*), pp. 746–51. Stroudsburg, PA: Assoc. Comput. Linguist.

Mitchell J, Lapata M. 2010. Composition in distributional models of semantics. *Cogn. Sci.* 34:1388–429

Padó S, Herbelot A, Kisselew M, Šnajder J. 2016. Predictability of distributional semantics in derivational word formation. In *Proceedings of the 26th International Conference on Computational Linguistics* (*COLING 2016*), pp. 1285–96. Stroudsburg, PA: Assoc. Comput. Linguist.

Pater J. 2019. Generative linguistics and neural networks at 60: foundation, friction, and fusion. *Language* 95:e41–74

Peters M, Neumann M, Iyyer M, Gardner M, Clark C, et al. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (*NAACL-HLT 2018*), pp. 2227–37. Stroudsburg, PA: Assoc. Comput. Linguist.

Pollard C, Sag IA. 1994. *Head-Driven Phrase Structure Grammar*. Chicago: Univ. Chicago Press

Pustejovsky J. 1995. *The Generative Lexicon*. Cambridge, MA: MIT Press

Reddy S, McCarthy D, Manandhar S. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pp. 210–21. Stroudsburg, PA: Assoc. Comput. Linguist.

Reisinger J, Mooney RJ. 2010. Multi-prototype vector-space models of word meaning. In *Proceedings of the 2010 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (*NAACL-HLT 2010*), pp. 109–17. Stroudsburg, PA: Assoc. Comput. Linguist.

Rosenfeld A, Erk K. 2018. Deep neural models of semantic shift. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (*NAACL-HLT 2018*), pp. 474–84. Stroudsburg, PA: Assoc. Comput. Linguist.

Rumelhart DE, Abrahamson AA. 1973. A model for analogical reasoning. *Cogn. Psychol.* 5:1–28

Sagi E, Kaufmann S, Clark B. 2009. Semantic Density Analysis: comparing word meaning across time and phonetic space. In *Proceedings of the EACL 2009 Workshop on GEMS: GEometical Models of Natural Language Semantics*, pp. 104–11. Stroudsburg, PA: Assoc. Comput. Linguist.

Sahlgren M. 2008. The distributional hypothesis. *Ital. J. Linguist.* 20:33–54

Santus E, Chersoni E, Lenci A, Blache P. 2017. Measuring thematic fit with distributional feature overlap. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (*EMNLP 2017*), pp. 648–58. Stroudsburg, PA: Assoc. Comput. Linguist.

Schulte im Walde S. 2006. Experiments on the automatic induction of {German} semantic verb classes. *Comput. Linguist.* 32:159–94

Schütze H. 1992. Dimensions of meaning. In *Proceedings of the 1992 ACM/IEEE Conference on Supercomputing*, pp. 787–96. Los Alamitos, CA: IEEE Comput. Soc.

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

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* (*EMNLP-CoNLL 2012*), pp. 1201–11. Stroudsburg, PA: Assoc. Comput. Linguist.

Springorum S, Schulte im Walde S, Utt J. 2013. Detecting polysemy in hard and soft cluster analyses of German preposition vector spaces. In *Proceedings of the International Joint Conference on Natural Language Processing*, pp. 632–40. Stroudsburg, PA: Assoc. Comput. Linguist.

Stern NG. 1921. *Swift, Swiftly and Their Synonyms: A Contribution to Semantic Analysis and Theory*. Gothenburg, Swed.: Wettergren & Kerber

Szymanski T. 2017. Temporal word analogies: identifying lexical replacement with diachronic word embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 448–53. Stroudsburg, PA: Assoc. Comput. Linguist.

Tahmasebi N, Borin L, Jatowt A. 2018. Survey of computational approaches to lexical semantic change. arXiv:1811.06278 [cs.CL]

Traugott EC, Dasher RB. 2001. *Regularity in Semantic Change*. Cambridge, UK: Cambridge Univ. Press

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

Vecchi EM, Marelli M, Zamparelli R, Baroni M. 2017. Spicy adjectives and nominal donkeys: capturing semantic deviance using compositionality in distributional spaces. *Cogn. Sci.* 41:102–36

Westera M, Boleda G. 2019. Don't blame distributional semantics if it can't do entailment. In *Proceedings of the 13th International Conference on Computational Semantics* (*IWCS 2019*), pp. 120–33. Stroudsburg, PA: Assoc. Comput. Linguist.

Xu Y, Kemp C. 2015. A computational evaluation of two laws of semantic change. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, pp. 2703–8. Austin, TX: Cogn. Sci. Soc.