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

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 multi-dimensional, graded, empirically induced word representations that successfully capture many aspects of meaning in natural languages, as shown in a large body of work 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 of relevance 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 review aims at fostering greater cross-fertilization of theoretical and computational approaches to language, as a means to advance our collective knowledge of how it works.

Contents

1. INTRODUCTION

This survey provides a critical discussion of the literature on distributional semantics, with an emphasis on methods and results that are of relevance 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, and subsequent work); yet, its impact in theoretical linguistics has so far been limited. A 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 survey. Accordingly, I will cover mostly research within computational linguistics, rather than cognitive science.

1.1. Distributional semantics in a nutshell

Here I provide only a brief introduction to distributional semantics, such that the survey is self-contained; for more comprehensive introductions, see Erk (2012), Clark (2015), 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 *post-doc* and *student*, are used in similar contexts (*a poor _, the _ struggled through the deadline*; examples from 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, illustrated in Figure 1, distributional semantics represents word meaning by taking large amounts of text as input and, through an abstraction mechanism (symbolized by the arrow), 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 multi-dimensional space (see below). However, many more possibilites are available and have been experimented with: The definition of distributional semantics encompasses all kinds of contexts, including for instance the visual context in 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).

Any grad student or **post-doc** he'd have would be a clonal copy of himself.
During that **post-doc**, I didn't publish much.
. . .

$$\Downarrow$$

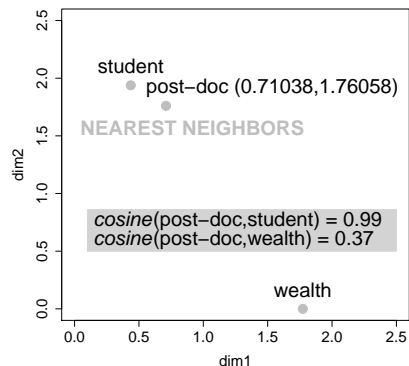|	dim1	dim2
post-doc	0.71038	1.76058
student	0.43679	1.93841
wealth	1.77337	0.00012

Figure 1: Distributional semantics in a nutshell: Inducing semantic representations from natural language data (left); visualizing and operating with these representations (right). Words are points in a space determined by the values in the dimensions of their vectors, like 0.71028 and 1.76058 for *post-doc*. *Post-doc* and *student* are nearer in semantic space than *post-doc* and *wealth*, and in fact they are nearest neighbors of (words closest to) each other. Adapted from Boleda & Herbelot (2016, Figure 1; CC-BY).

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, and the graph in Figure 1 (right) is a graphical rendering of our toy lexicon. The vectors for *post-doc* and *student* are closer in the space than those of *post-doc* 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, the larger the cosine similarity. For instance, the cosine between *post-doc* and *student* in our space is 0.99, while it is 0.37 for *post-doc* vs. *wealth* (cosine values for positive vectors range between 0 and 1).

Our example is two-dimensional, but in actual distributional models many more dimensions are used, 300-400 being a frequent range. While we cannot represent so many dimensions visually, the geometric properties of two-dimensional spaces that we discuss 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 *post-doc* in our mini semantic space.

Finally, there are many different versions of the abstraction function (the arrow in Figure 1). Earlier distributional models were built by extracting and transforming co-

occurrence statistics, while recently models based on neural networks have gained ground due to their good performance (Baroni et al. 2014b). Neural networks are a versatile machine learning type of algorithm, used for tasks like machine translation or image labeling; for reasons of scope, in this survey we 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 largely arises from structuralist traditions (Sahlgren 2008). As in structuralism, words are defined according to their position in a system, the lexicon, based on 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 like ±MALE are instead captured in a distributed fashion, as varying patterns across the whole vector. There are three further key differences to traditional feature-based approaches in linguistics that render distributional semantics attractive as a model of word meaning.

First, the fact that distributional representations are <u>learnt</u> from natural language data, and thus 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. This provides semantic representations on a large scale, in a single, coherent system where systematic explorations are possible.

Table 1: Near-synonyms in semantic space: The words closest to *man*, *chap*, *lad*, and *guy* in the distributional model of Baroni et al. (2014b), adapted from Baroni (2016a).

Word	Nearest neighbors
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

Second, high <u>multi-dimensionality</u>. The information abstracted from the data is distributed across all the dimensions of a vector, typically a few hundred, which allows for 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/lad/chap/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, like ±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, like *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*).

Third, and relatedly, 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

Diachronic semantics, especially lexical semantic change, is an area where the interaction between use and meaning (crucial for distributional semantics) has traditionally been the focus of interest already in theoretical linguistics (Traugott & Dasher 2001; Deo 2015). For instance, the word *gay* gradually changed during the 20th century from a meaning similar to 'cheerful' to its current predominant use as 'homosexual', and its contexts of use in language reflect this change: Examples in (1) are from the year 1900, and those in (2) from 2000 (source: COHA, Davies 2010-). 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 it 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 we will see below, its data-driven nature allows it to detect semantic change from changes in usage.

(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.

2.1. Distributional approaches to diachronic semantics

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

This is typically done 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 Book Ngrams corpus (Michel et al. 2011). The cosine similarity of the word *gay*, when compared to its representation in the 1900 lexicon,

goes down through the 20th century, with the drop accelerating at the end of the 70s from around 0.75 to around 0.3 in 2000.

Figure 2 visualizes the trajectory of three words across time in another study (Hamilton et al. 2016), with nearest neighbors in gray font along the words of interest. It illustrates how inspection of nearest neighbors can help trace the specific meaning shift taking place. In 1900, *gay* is near words like *daft* or *cheerful*, and by 1990 it is instead near to *homosexual*. the change in *broadcast* is metaphorical in nature, from a concrete to a more abstract meaning (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 is targeted comparisons to words related to the old and the new meanings: For instance, Kim et al. (2014) compare how the cosine similarities of *cell* to *dungeon* and *phone* evolve through the years. However, the latter requires previous knowledge of the specific change taking place.
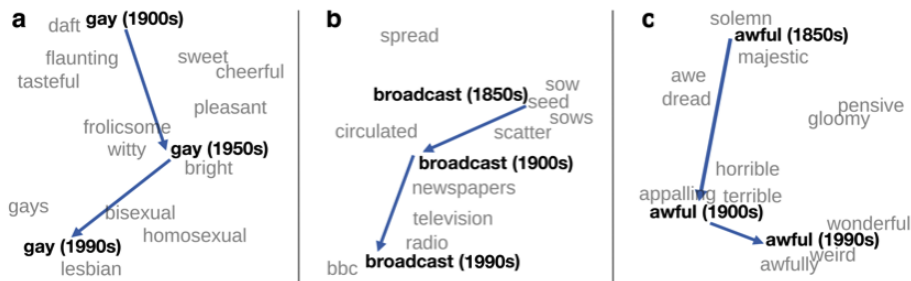

Figure 2: Two-dimensional visualization of semantic change for three English words, reproduced from Hamilton et al. (2016). The figure was obtained via dimensionality reduction from a space with 300-dimensional vectors. Reproduced with permission.

There is current experimentation on two related efforts (Tahmasebi et al. 2018): sense-specific semantic change, where sense representations are induced and then tracked (also see Section 3.2), and detecting not only the presence but also the type of semantic shift. In the latter literature, starting with the pioneering work of Sagi et al. (2009), there is 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; one 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 verb *do* going from a lexical to an auxiliary verb between the 15th and the 18th century. Sagi et al. (2009) trace these three words by representing each context of use individually, with one vector per sentence. They show that, for *dog* and *do*, contexts become more separate over time, corresponding to the broadening effect, and the reverse 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.

Up to now, 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 Xu & Kemp (2015), a study assessing two previously proposed laws that make contradicting predictions. Their large-scale computa-

tional analysis, based on distributional semantic models of English on the Google Ngrams corpus, shows that pairs of synonyms tend to stay closer in space than control pairs across the 20th century, in four dataset jointly comprising tens of thousands of words. They thus provide support for the law of parallel change (Stern 1921), that posits that related words undergo similar changes, and against the law of differentiation (Bréal 1897), that defends that synonyms tend to evolve different meanings because it is not efficient for languages to maintain synonyms. Other generalizations about semantic change emerging from work 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 points: (1) detecting semantic change, as a change in the representation of a word across time; (2) temporally locating it, by monitoring the rate of change in the distributional representation; (3) tracking the specific semantic evolution of the word, via an inspection of the nearest neighbors or targeted examination of cosine similarities; (4) testing competing theories via large-scale empirical studies. It can also help detecting the type of semantic change, although this is still an under-researched topic.

A major challenge is the fact that distributional methods, especially those based on neural networks, are quite data-hungry, while many datasets 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 datasets used for diachronic semantics, identified in Tahmasebi et al. (2018)'s survey, only 4 are not in English. Moreover, the vast majority of studies focus on the Google Book Ngrams corpus, which covers only the 1850-2009 period.

When the amount of data is scarce, spurious effects easily arise. For instance, Del Tredici et al. (2019), in a study of meaning shift in a community of soccer fans with data from 2011 to 2017, 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 mostly used when talking about a new coach.[1] Effects like this 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 time scales, because of the larger amount 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 mostly focused on lexical semantic change, while in diachronic semantics there is 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 do provide usable representations for them (Mikolov et al. 2010; Peters et al. 2018), opening new possibilities.

---

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

3. POLYSEMY AND COMPOSITION

Words are notoriously ambiguous or polysemous, that is, they adopt different meanings in different contexts (Cruse 1986, among many others). For instance, *post-doc* refers to a person in the first sentence in Figure 1, and to a period of time in the second. Distributional semantics has traditionally tackled this issue in two ways, which resonate with linguistic treatments of polysemy (Lyons 1977). The predominant approach, by far, is to take the word as a unit of representation and provide a single representation that encompasses all 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 like the Generative Lexicon (Pustejovsky 1995), Frame Semantics (Fillmore et al. 2006), or HPSG (Pollard & Sag 1994), even if not directly inspired by them. These approaches aim at encoding 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, 122-123) 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, the change of state sense emerges; if it combines with an artifact, the creation sense emerges. This 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 (see Pustejovsky 1995; Kilgarriff 1997, for discussion).

corpus-based

	dim1	dim2
cut	4	5
cost	1	5
cut cost	4	9

synthetic

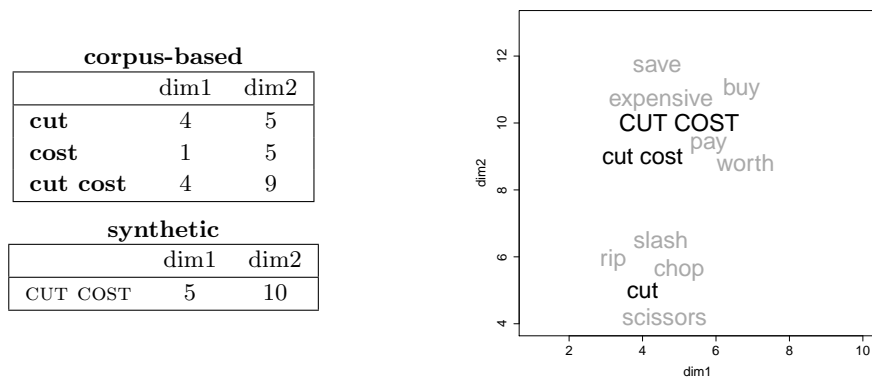|	dim1	dim2
CUT COST	5	10

Figure 3: Compositional distributional semantics: Illustration with vector addition. Left: The synthetic vector CUT COST is built by component-wise addition of the vectors for *cut* and *cost*. Right: The argument *cost* pulls the vector for *cut* towards 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.

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). The pioneering work of Kintsch (2001) in Cognitive Science started extending distributional methods to compose such generic word representations into larger constituents. The computational linguistic community took this research line up almost a decade later (Erk & Padó 2008; Mitchell & Lapata 2010; Baroni & Zamparelli 2010; Coecke et al. 2011; Socher et al. 2012; Mikolov et al. 2013a). Compositional distributional methods build representations for phrases out of the representations of their parts, and the corresponding meaning is expected to emerge as a result of the composition (Baroni et al. 2014a). Figure 3 provides an illustration, with 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 (*dim1*) because the values of *cut* and *cost* for this dimension are 4 and 1, respectively.

To give an intuition of how this may account for semantic effects, let's assume that dimension 1 is associated to abstract notions and dimension 2 to concrete notions (of course this is a simplification; remember that properties like 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*, and so it has high values for both dimensions. Instead, *cost* is an abstract notion, and so it has low values for dimension 1 and high values for dimension 2. When composing the two, the abstract dimension gets highlighted, pulling the vector towards regions in the semantic space related to its abstract sense. This is shown in Figure 3 (right); while the vector values are fictitious, the neighbors (in gray) 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 towards the abstract sense. This mechanism by which matching semantic dimensions reinforce each other, and mismatched dimensions remain less active, is reminiscent of the mechanisms by Pustejovsky (1995) discussed above for *bake a potato* vs. *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; rather, the structure is implicitly defined in the space.

A substantial body of work has shown that composition methods in distributional semantics largely account for polysemy effects in semantic composition. Baroni & Zamparelli (2010) and subsequent work compare the synthetic vector for a phrase, like 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. In 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 average cosine similarity of 0.4 to their head nouns. Another common method is to compare model results to human intuitions about the semantics of phrases. Mitchell & Lapata (2010) introduced this for phrase similarity (in turn inspired on methods to evaluate word similarity), with participant data such as *reduce amount - cut cost* being very similar, *encourage child - leave company* 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 work experiments with

---

[2]User-friendly interface to this semantic space: `http://meshugga.ugent.be/snaut-english`.

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 not much work on function words because, as mentioned above, these are traditionally hard to model with distributional semantics. An exception is Bernardi et al. (2013), which seeks 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 are able to distinguish between semantically acceptable and unacceptable adjective-noun phrases. Crucially, their data involves phrases that are unattested in a very large corpus; some phrases are unattested because they are semantically anomalous (*angry lamp*, *legislative onion*), and some due to 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 (*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.

Work in this area has investigated much more sophisticated approaches to composition than vector addition. I cannot do justice to this research for reasons of space, but see Erk (2012) and Baroni (2013). Much of this work is inspired by formal semantics (Garrette et al. 2011; Beltagy et al. 2013; Baroni et al. 2014a; Coecke et al. 2011; Lewis & Steedman 2013; Herbelot & Vecchi 2015; Erk 2016); Boleda & Herbelot (2016) surveys 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 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 remains 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, like *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 like 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, cf. 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 seen; the results in 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 semi-opacity, which is the opposite end of compositionality; see Section 4.2 for work on the compositionality of noun compounds.

3.2. Different representations, polysemy via word senses

Other work aims at building 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 each other than to uses of the same word in a different sense, we can detect word senses by checking the similarity of the contexts. The pioneering work of Schütze (1998) did this 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. This work started a tradition, within distributional semantics, of research on word sense induction and sense-specific word representations (McCarthy et al. 2004; Reisinger & Mooney 2010). Erk and colleagues instead aimed at providing a representation of the specific meaning a word takes in a given context, entirely bypassing word senses (Erk & Padó 2008; Erk & Padó 2010; Erk et al. 2013). This work 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 1) deciding when two senses are different enough to warrant the addition of an item to the vocabulary, 2) 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 (like the abstract/concrete axis of *cut*), in a graded fashion. The same way that it 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). Multi-dimensionality allows it to capture both the common core to 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 (like ALLEGED KILLER) will be further away from their corpus-based vectors than synthetic phrases with other adjectives (like 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, like *severe* for *pain*, then it is easy to model; if it is less typical, like *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). The authors propose that composition exploits two aspects of meaning: on the one hand, the conceptual aspect, with regularities in how words match (*box* denotes a physical object, and *red* is typically used to specify colors of physical objects); on the other hand, the referential aspect, specifically the information about the referent of the phrase (for instance, *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) took distributional semantic data themselves as an object of empirical inquiry; they asked what made certain phrases difficult for compositional distributional models, and the answer proved theoretically worthy. This work is thus 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 for instance in 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 work 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, and thus it is natural to turn it around and use distributional cues to infer semantic classes —as Levin herself does in her research in a manual fashion.

Levin's work had a big impact in Computational Linguistics, inspiring work on the automatic acquisition of semantic classes from distributional evidence (Dorr & Jones 1996; Merlo & Stevenson 2001; McCarthy 2000; Korhonen et al. 2003; Lapata & Brew 2004; Schulte im Walde 2006; Boleda et al. 2012). 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 work targets a finer-grained classification, with Levin-style semantic classes, such as Schulte im Walde (2006) for German. This early work used distributional evidence, but not distributional semantics strictu sensu. Baroni & Lenci (2010) replicated Merlo & Stevenson (2001)'s experiment using a proper distributional model, obtaining comparable accuracies.

Erk et al. (2010) initiated a line of work using distributional methods to model selectional restrictions, or the thematic fit between an argument and a predicate (usually, a verb). They capitalize 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 is in turn better than *sympathy*. Again, this 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*, etc. (see Figure 4). This average vector can be seen as a prototype for the argument of the verb.
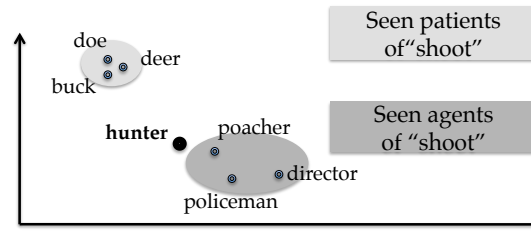

Figure 4: Visualization of the approach to selectional preferences in Erk et al. (2010). Partial reproduction of their Figure 1, p. 731; CC-BY.

Erk et al. compared the scores of the model to human ratings (where participants were asked to rate the plausibility that e.g. *hunter* is an agent of *shoot*). Their model achieved a Spearman correlation score of 0.33 and 0.47 ($p < 0.001$) with the human ratings in two different datasets 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; Lenci 2011; Greenberg et al. 2015; 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 have some agentive connotation, like *carver*). Derivational processes are semi-regular; they are largely compositional, but not always (mainly due to lexicalization processes), and they present subregularities (for instance, *carver, driver* denote agents, but *broiler, cutter* denote instruments). Moreover, both stem and affix semantics exhibit the properties typical of word semantics, such as polysemy and gradeness (Marelli & Baroni 2015); cf. the polysemy of *-er* between agent and instrument. Thus, accounting for morphological derivation requires fine-grained lexical semantic representations for both stem and affix and 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 (Lazaridou et al. 2013; Marelli & Baroni 2015; Padó et al. 2016; Lapesa et al. 2018; Cotterell & Schütze 2018, a.o.); most of this work has adapted compositional methods initially developed for

word composition. The motivation in this work is two-fold. 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, and thus 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 is simply unavailable, making it 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 up 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 to obtain affix representations is to average derived words (Padó et al. 2016) —for instance, averaging the vectors for *carver*, *drinker*, *driver*, etc. to obtain a representation for *-er*.

Table 2: Derivational phenomena captured with compositional distributional semantic methods: Examples from Marelli & Baroni (2015).

Phenomenon	Word	Nearest neighbors (selection)[a]
Affix polysemy	CARVER[b]	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, hard-working

[a]Marelli & Baroni (2015) provide a selection of the 20 nearest neighbors; [b]in small caps, synthetic word representations, produced by derivation operations with distributional semantics (see text).

Table 2 showcases phenomena captured by the distributional model of Marelli & Baroni (2015), illustrated through nearest neighbors (see the 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). Words in small caps correspond to synthetic word vectors, the rest to corpus-based word vectors. The first block 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, like *potter* or *goldsmith*, whereas BROILER is in the region of cooking instruments (*oven*, *stove*). In the second block, we see 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 nearest neighbors), but *-ist* correctly focuses on the sense related to 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 block 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 they are able to do so is still under investigation, since distributional methods exhibit a large variance in performance across invidual words and across derivational patterns (Padó et al. 2016). The factors intervening 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 unfrequent, 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. It is thus 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 to *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 alleviate the data problem explained above. For instance, the nearest neighbors of *rename* in their distributional space are *defunct, officially, merge*, whereas those for the synthetic vector RENAME are *name, later, namesake.*

As for linguistic factors, for instance Padó et al. (2016), in a large-scale study of derivation in German, find that the derivational pattern is the best predictor for model performance (that is, 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 was the case with composition (see Section 3.1). This again suggests that the distributional representations themselves do most of the job, and are more important than the specific method used. Cotterell & Schütze (2018)'s results underscore this interpretation. They 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 emerged also in the study of semantic opacity and semi-opacity, which typically aims at predicting the degree of compositionality in compound nouns and multi-word expressions. Reddy et al. (2011), a representative study, aimed at reproducing human ratings on the degree of compositionality of 90 English compound nouns (*climate change, graduate student, speed limit* obtaining maximum compositionality scores, *silver bullet, ivory tower, gravy train* minimum). Adding the two component vectors (with a higher weight of the modifier; see paper for details) achieved a Spearman correlation score of 0.71 with human data. Other work uses different methods; for instance, Springorum et al. (2013) do not use compositional methods but explore how the modifier and the head contribute to compositionality ratings for German data. Against their prediction, the

modifier is a much better predictor of compositionality than the head.

Again, most work is directed at showing <u>that</u> distributional semantics can model derivational morphology, rather than tackling more specific linguistic questions. An exception is 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 predictions is that diminutives shift words towards positive valence (consider *Hund* → *Hündchen*, English *dog-doggie*). The work 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*, Eng. *drive-run over* vs. *nehmen* → *übernehmen*, Eng. *take-take over*). This work again showcases 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 Mikolov et al. (2013b), which 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 you subtract an inflected word from its stem, you obtain 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. 40% may not look 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 work 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 (Springorum et al. 2013; Lazaridou et al. 2013) or 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 to the average of all the bases in the pattern (e.g. for *-er*, the word vector for *carve* compared to the average of the vectors for *carve, drink, drive, broil, cut*, etc.).

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

5. CONCLUSION AND OUTLOOK

The above discussion summarizes robust results in distributional semantics that can be directly imported for research in theoretical linguistics, as well as challenges and open issues. Among the robust results are that (1) distributional semantics is particularly useful in areas where the connection between use, meaning, and grammar is relevant, such as the areas reviewed in this survey; (2) geometric relationships in distributional models correspond to semantic relationships in language; (3) gradedness in distributional representations correlates with gradedness in semantic phenomena (e.g., the degree of semantic change); (4) averaging the distributional representations of classes of words yields useful abstractions of the relevant classes (e.g., of arguments accepted by specific predicates); (5) simple combinations of distributional representations produce quite accurate predictions as to the semantics of phrases and derived words. I have argued that the multi-dimensional, graded, and data-driven nature of its representation 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 testbed 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. This requires collaboration between computational and theoretical linguists.

There are also a number of challenges that distributional methods face. Like other data-driven methods, distributional models mirror the data they are fed. This is good, because they provide radically empirical representations, and also dangerous, because representations are subject to biases in the underlying data (Caliskan et al. 2017). A related challenge is the fact that distributional methods need large amounts 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 for the moment. A final, crucial issue is the lack of adequate researcher training, which prevents a wider use of distributional semantics in linguistics, and of quantitative and computational methods more generally. Strengthening student training in such methods in linguistics degrees is of paramount importance to allow the field to adequately exploit the vast amounts of linguistic data that have become available in the last few decades.

In this survey, 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 are 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 in the survey for more information.

For reasons of scope, I have also left out of the discussion 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, like 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 said in this survey applies to such modules. However, crucially, these systems also have other modules that represent linguistic context, and 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, the examples in (3)-(4) (from 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 (3), referring to signing for plays rather than to the acting itself in (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 [. . .]
         NN: [. . .] they were actors who had been handed fat roles in a successful play [. . .]

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

DISCLOSURE STATEMENT

ACKNOWLEDGMENTS

LITERATURE CITED

Aina L, Gulordava K, Boleda G. 2019. Putting words in context: LSTM language models and lexical ambiguity, In *Proc. of ACL*. Florence, Italy: ACL

Alishahi A, Chrupała G, Linzen T. 2019. Analyzing and interpreting neural networks for NLP: A report on the first BlackboxNLP Workshop. *arXiv preprint arXiv:1904.04063*

Arora S, Li Y, Liang Y, Ma T, Risteski A. 2018. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics* 6:483–495

Baayen RH, Piepenbrock R, van H R. 1993. The CELEX lexical data base on CD-ROM. Linguistic Data Consortium

Baroni M. 2013. Composition in distributional semantics. *Language and Linguistics Compass* 7:511–522

Baroni M. 2016a. Composes: An executive summary. Talk at the Composes Workshop

Baroni M. 2016b. Grounding Distributional Semantics in the Visual World. *Linguistics and Language Compass* 10:3–13

Baroni M, Bernardi R, Zamparelli R. 2014a. Frege in space: A program for compositional distributional semantics. *Linguistic Issues in Language Technology* 9:5–110

Baroni M, Dinu G, Kruszewski G. 2014b. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors, In *Proc. of ACL*. Baltimore, Maryland: ACL

Baroni M, Lenci A. 2010. Distributional Memory: A General Framework for Corpus-Based Semantics. *Computational Linguistics* 36:673–721

Baroni M, Zamparelli R. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space, In *Proc. of EMNLP*. ACL

Beltagy I, Chau C, Boleda G, Garrette D, Erk K, Mooney R. 2013. Montague meets Markov: Deep semantics with probabilistic logical form, In *Proceedings of Starsem*

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. *Language Resources and Evaluation* 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 *Proc. of ACL*. Sofia, Bulgaria: ACL

Bojanowski P, Grave E, Joulin A, Mikolov T. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5:135146

Boleda G, Baroni M, Pham TN, McNally L. 2013. Intensionality was only alleged: On adjective-noun composition in distributional semantics, In *Proc. of IWCS*. Potsdam, Germany: ACL

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

Boleda G, Herbelot A. 2016. Formal distributional semantics: Introduction to the special issue. *Computational Linguistics* 42

Boleda G, Schulte im Walde S, Badia T. 2012. Modeling regular polysemy: A study on the semantic classification of catalan adjectives. *Computational Linguistics* 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* 356:183–186

Camacho-Collados J, Pilehvar MT. 2018. From word to sense embeddings: A survey on vector

representations of meaning. *Journal of Artificial Intelligence Research* 63:743–788

Clark S. 2015. Vector space models of lexical meaning. *The Handbook of Contemporary semantic theory* :493–522

Coecke B, Sadrzadeh M, Clark S. 2011. Mathematical Foundations for a Compositional Distributional Model of Meaning. *Linguistic Analysis: A Festschrift for Joachim Lambek* 36:345–384

Cotterell R, Schütze H. 2018. Joint semantic synthesis and morphological analysis of the derived word. *Transactions of the Association for Computational Linguistics* 6:33–48

Cruse DA. 1986. Lexical semantics. Cambridge, UK: Cambridge University Press

Davies M. 2010-. The Corpus of Historical American English (COHA): 400 million words, 1810-2009

Del Tredici M, Fernández R, Gemma B. 2019. Short-term meaning shift: a distributional exploration, In *Proc. of NAACL-HLT*. Denver, Colorado: ACL

Deo A. 2015. Diachronic semantics. *Annu. Rev. Linguist.* 1:179–197

Dorr BJ, Jones D. 1996. Role of Word Sense Disambiguation in Lexical Acquisition: Predicting Semantics from Syntactic Cues, In *Proc. of COLING-96*. Morristown, NJ: ACL

Dubossarsky H, Weinshall D, Grossman E. 2017. Outta control: Laws of semantic change and inherent biases in word representation models, In *Proc. of EMNLP*. Copenhagen, Denmark: ACL

Erk K. 2012. Vector Space Models of Word Meaning and Phrase Meaning: A Survey. *Linguistics and Language Compass* 6:635–653

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

Erk K, McCarthy D, Gaylord N. 2013. Measuring Word Meaning in Context. *Computational Linguistics* 39:511–554

Erk K, Padó S. 2008. A structured vector space model for word meaning in context, In *Proc. of the 2008 EMNLP*. Honolulu, Hawaii: ACL

Erk K, Padó S. 2010. Exemplar-based models for word meaning in context, In *Proc. of ACL*. Stroudsburg, PA, USA: ACL

Erk K, Padó S, Padó U. 2010. A Flexible, Corpus-driven Model of Regular and Inverse Selectional Preferences. *Computational Linguistics* 36:723–763

Fillmore CJ, et al. 2006. Frame semantics. *Cognitive linguistics: Basic readings* 34:373–400

Garrette D, Erk K, Mooney R. 2011. Integrating logical representations with probabilistic information using Markov logic, In *Proc. of IWCS*. ACL

Greenberg C, Sayeed A, Demberg V. 2015. Improving unsupervised vector-space thematic fit evaluation via role-filler prototype clustering, In *Proc. NAACL-HLT*. Denver, Colorado: ACL

Grefenstette E, Sadrzadeh M. 2011. Experimenting with transitive verbs in a DisCoCat, In *Proc. of the GEMS 2011 Workshop*. Edinburgh, UK: ACL

Grimshaw J. 1990. Argument structure. the MIT Press

Gulordava K, Baroni M. 2011. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus, In *Proc. of the GEMS 2011 Workshop*. ACL

Hamilton WL, Leskovec J, Jurafsky D. 2016. Diachronic word embeddings reveal statistical laws of semantic change, In *Proc. of ACL*

Harris ZS. 1954. Distributional structure. *Word* 10:146–162

Herbelot A, Vecchi EM. 2015. Building a shared world: Mapping distributional to model-theoretic semantic spaces, In *Proc. of EMNLP*. Lisbon, Portugal: ACL

Heylen K, Wielfaert T, Speelman D, Geeraerts D. 2015. Monitoring polysemy: Word space models as a tool for large-scale lexical semantic analysis. *Lingua* 157:153–172

Hock HH. 1991. Principles of historical linguistics. Walter de Gruyter

Kilgarriff A. 1997. I don't believe in word senses. *Computers and the Humanities* 31:91–113

Kim Y, Chiu YI, Hanaki K, Hegde D, Petrov S. 2014. Temporal analysis of language through neural language models, In *Proc. of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. ACL

Kintsch W. 2001. Predication. *Cognitive science* 25:173–202

Korhonen A, Krymolowski Y, Marx Z. 2003. Clustering polysemic subcategorization frame distributions semantically, In *Proc. of ACL*. Stroudsburg, PA, USA: ACL

Kutuzov A, Øvrelid L, Szymanski T, Velldal E. 2018. Diachronic word embeddings and semantic shifts: a survey, In *Proc. of COLING*. ACL

Landauer TK, Dumais ST. 1997. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological review* 104:211–240

Lapata M, Brew C. 2004. Verb Class Disambiguation Using Informative Priors. *Computational Linguistics* 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 Structure* 11:277–312

Lapesa G, Pado S, Pross T, Rossdeutscher A. 2017. Are doggies really nicer than dogs? the impact of morphological derivation on emotional valence in German, In *Proc. of IWCS*. ACL

Lazaridou A, Marelli M, Zamparelli R, Baroni M. 2013. Compositional-ly Derived Representations of Morphologically Complex Words in Distributional Semantics, In *Proc. of ACL*. Sofia, Bulgaria: ACL

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

Lenci A. 2011. Composing and Updating Verb Argument Expectations: A Distributional Semantic Model, In *Proc. of the 2nd Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*. ACL

Lenci A. 2018. Distributional models of word meaning. *Annual review of Linguistics* 4:151–171

Levin B. 1993. English verb classes and alternations: A preliminary investigation. Chicago and London: University of Chicago Press

Lewis M, Steedman M. 2013. Combined Distributional and Logical Semantics. *Transactions of the Association for Computational Linguistics* 1:179–192

Lieber R. 2004. Morphology and lexical semantics. Cambridge University Press

Lyons J. 1977. Semantics. Cambridge University 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. *Journal of Memory and Language* 92:57–78

Marelli M, Baroni M. 2015. Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological Review* 122:485–515

McCarthy D. 2000. Using semantic preferences to identify verbal participation in role switching alternations, In *Proc. of NAACL*. San Francisco, USA: Morgan Kaufmann Publishers Inc.

McCarthy D, Koeling R, Weeds J, Carroll J. 2004. Finding predominant word senses in untagged text, In *Proc. of ACL*. Barcelona, Spain

McNally L, Boleda G. 2017. Conceptual vs. Referential Affordance in Concept Composition. In *Compositionality and concepts in linguistics and psychology*, eds. JA Hampton, Y Winter, Language, Cognition, and Mind. Springer International Publishing, 245–267

Merlo P, Stevenson S. 2001. Automatic Verb Classification Based on Statistical Distributions of Argument Structure. *Computational Linguistics* 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* 331:176–182

Mikolov T, Karafiat M, Burget L, Cernocky J, Khudanpur S. 2010. Recurrent Neural Network based Language Model, In *Proc. of Interspeech*

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. 2013a. Distributed representations of words and phrases and their compositionality, In *Advances in neural information processing systems*

Mikolov T, Yih Wt, Zweig G. 2013b. Linguistic Regularities in Continuous Space Word Representations, In *Proc. of NAACL-HLT*. Atlanta: ACL

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

Padó S, Herbelot A, Kisselew M, Šnajder J. 2016. Predictability of distributional semantics in derivational word formation, In *Proc. of COLING*. Osaka, Japan: The COLING 2016 Organizing Committee

Pater J. 2019. Generative linguistics and neural networks at 60: Foundation, friction, and fusion. *Language* 95:e41–e74

Peters M, Neumann M, Iyyer M, Gardner M, Clark C, et al. 2018. Deep contextualized word representations, In *Proc. of NAACL-HLT*. New Orleans, Louisiana: ACL

Pollard C, Sag IA. 1994. Head-driven phrase structure grammar. University of Chicago Press

Pustejovsky J. 1995. The Generative Lexicon. Cambridge, MA (etc.): The MIT Press

Reddy S, McCarthy D, Manandhar S. 2011. An empirical study on compositionality in compound nouns, In *Proc. of IJCNLP*. ACL

Reisinger J, Mooney RJ. 2010. Multi-prototype vector-space models of word meaning, In *Proc. of NAACL-HLT*. Los Angeles, California: ACL

Rosenfeld A, Erk K. 2018. Deep neural models of semantic shift, In *Proc. of NAACL-HLT*. New Orleans, Louisiana: ACL

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

Sagi E, Kaufmann S, Clark B. 2009. Semantic Density Analysis: Comparing word meaning across time and phonetic space, In *Proc. of the EACL 2009 GEMS Workshop*. ACL

Sahlgren M. 2008. The distributional hypothesis. *Italian journal of linguistics* 20:33–54

Santus E, Chersoni E, Lenci A, Blache P. 2017. Measuring thematic fit with distributional feature overlap, In *Proc. of EMNLP*. Copenhagen, Denmark: ACL

Schulte im Walde S. 2006. Experiments on the Automatic Induction of {German} Semantic Verb Classes. *Computational Linguistics* 32:159–194

Schütze H. 1992. Dimensions of meaning, In *Proc. of Supercomputing '92*. Minneapolis, USA

Schütze H. 1998. Automatic word sense discrimination. *Computational Linguistics* 24:97–123

Socher R, Huval B, Manning CD, Ng AY. 2012. Semantic compositionality through recursive matrix-vector spaces, In *Proc. of EMNLP-CoNLL*. Jeju Island, Korea: ACL

Springorum S, Schulte im Walde S, Utt J. 2013. Detecting polysemy in hard and soft cluster analyses of German preposition vector spaces, In *Proc. of IJCNLP*. Nagoya, Japan: Asian Federation of Natural Language Processing

Stern NG. 1921. Swift, swiftly and their synonyms. A contribution to semantic analysis and theory. Gothenburg: Wettergren & Kerber

Szymanski T. 2017. Temporal word analogies: Identifying lexical replacement with diachronic word embeddings, In *Proc. of ACL*. Vancouver, Canada: ACL

Tahmasebi N, Borin L, Jatowt A. 2018. Survey of Computational Approaches to Lexical Semantic Change. *arXiv preprint arXiv:1811.06278*

Traugott EC, Dasher RB. 2001. Regularity in semantic change. Cambridge University Press

Turney PD, Pantel P. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research* 37:141–188

Vecchi EM, Marelli M, Zamparelli R, Baroni M. 2017. Spicy Adjectives and Nominal Donkeys: Capturing Semantic Deviance Using Compositionality in Distributional Spaces. *Cognitive Science* 41:102–136

Westera M, Boleda G. 2019. Don't blame distributional semantics if it can't do entailment, In *Proc. of IWCS*. Gothenburg, Sweden: ACL

Xu Y, Kemp C. 2015. A computational evaluation of two laws of semantic change, In *Proc. of CogSci*. Pasadena, California, USA