Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions

Fritz Günther[1], Luca Rinaldi[1,2], & Marco Marelli[1,2]
[1]University of Milano–Bicocca, Milan, Italy
[2]NeuroMI, Milan Center for Neuroscience, Milan, Italy

Models representing meaning as high-dimensional numerical vectors (such as LSA, HAL, BEAGLE, Topic Models, GloVe or word2vec) have been introduced as extremely powerful machine-learning proxies for human semantic representations, and have seen an explosive rise in popularity over the last two decades. However, despite their considerable advancements and spread in the cognitive sciences, one can observe problems associated with the adequate presentation and understanding of some of their features. Indeed, when these models are examined from a cognitive perspective, a number of unfounded arguments tend to appear in the psychological literature. In the present article, we review the most common of these arguments, directed at (1) what exactly these models represent at the implementational level and their plausibility as a cognitive theory, (2) how they deal with various aspects of meaning such as polysemy or compositionality, and (3) how they relate to the debate on embodied and grounded cognition. We identify common misconceptions arising due to incomplete descriptions, outdated arguments, and unclear distinctions between theory and implementation of the models. We clarify and amend these points, to provide a theoretical basis for future research and discussions on vector models of semantic representation.

Keywords: Distributional Semantic Models; Semantic Representations; Latent Semantic Analysis; Computational Models of Meaning; Semantic Memory

This is the author's pre-print version of the article (date: 08. 05. 2019), to be published in *Perspectives on Psychological Science*.

Computationally implemented theories of human semantic representations, in which word meanings are represented as high-dimensional numerical vectors extracted from large amounts of natural language data, appeared in the field of cognitive science about twenty years ago. The most prominent early models in the field were LSA (Latent Semantic Analysis; Landauer & Dumais, 1997) and HAL (Hyperspace Analogue to Language; Lund & Burgess, 1996). These models and their successors, such as BEAGLE (Bound Encoding of the AGgregate Language Environment; Jones & Mewhort, 2007), Topic Models (Griffiths, Steyvers, & Tenenbaum, 2007) or, more recently, GloVe (Global Vectors; Pennington, Socher, & Manning, 2014) and *word2vec* (Mikolov, Chen, Corrado, & Dean, 2013; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) have since then received considerable attention, both as applied models for word meaning induction and as the focus of investigation in cognitive science. An overview of these models, alongside with short descriptions, is provided in Table 1.

These word vector models have been shown to be impressively high-performing: For example, they have achieved perfect scores on multiple choice tests (Bullinaria & Levy, 2012), above 95 % purity in word categorization (Baroni & Lenci, 2010), and correlations around .8 with human word similarity ratings – a score comparable to human inter-rater agreements (Baroni, Dinu, & Kruszewski, 2014; Bruni, Tran, & Baroni, 2014). Therefore, they have been successfully applied in virtually all fields of cognitive science, including artificial intelligence research (see Tur-

This work was supported by a Research Fellowship (no. 392225719) from the German Research Foundation (DFG), awarded to Fritz Günther, and by grant 2017-1633 from the Fondazione Cariplo–Regione Lombardia, awarded to Marco Marelli. We thank the participants and speakers of the CARLA 2018 (Osnabrück) and SINP 2017 (Palermo) for many helpful and insightful discussions which inspired many sections of this article. We also thank Anna Borghi and Art Glenberg for many valuable comments and discussions on earlier versions of this manuscript. Finally, we thank Marc Brysbaert, Mike Jones, and two anonymous reviewers for their very constructive feedback to earlier versions of this manuscript, and Laura A. King for her support in the editorial process.

Table 1
The most prominent vector models of semantic representation, along with short descriptions. More detailed descriptions are provided in later sections of the article.

Model	Authors	Year	Venue of Publication	Short Description
HAL	Lund & Burgess	1996	*Behavior Research Methods*	Creates a Word-by-Word Matrix
LSA	Landauer & Dumais	1997	*Psychological Review*	Creates a Word-by-Document Matrix and applies dimensionality reduction via SVD
Topic Models	Griffiths, Steyvers & Tenenbaum	2007	*Psychological Review*	Creates a Word-by-Document Matrix and applies dimensionality reduction via LDA
BEAGLE	Jones & Mewhort	2007	*Psychological Review*	Modifies initially random word vectors based on the other words in successively processed documents
word2vec	Mikolov, Chen, et al.; Mikolov, Sutskever, et al.	2013	*ICLR Workshop*, *Neural Information Processing Systems*	Trains word vectors as the hidden layer of a neural network that predicts words from the surrounding words, or vice versa
GloVe	Pennington, Socher & Manning	2014	*Empirical Methods in Natural Language Processing*	Trains word vectors to optimally predict words' probability of co-occurrence

ney & Pantel, 2010), computational psychology (see Jones, Willits, & Dennis, 2015), psycholinguistics (e.g., Jones, Kintsch, & Mewhort, 2006), cognitive neuroscience (e.g., T. M. Mitchell et al., 2008), instructional design and education (e.g. Wade-Stein & Kintsch, 2004), but also social psychology (e.g. Lenton, Sedikides, & Bruder, 2009), psychiatry (e.g. Elvevåg, Foltz, Weinberger, & Goldberg, 2007) and biomedicine (e.g. Cohen & Widdows, 2009).

Accordingly, these models have been highly impactful: As of the time of this writing (April 2019), the article by Landauer and Dumais (1997) on the Latent Semantic Analysis (LSA) model has been cited almost 6,400 times on Google Scholar, the article by Lund and Burgess (1996) on the Hyperspace Analogue to Language (HAL) model about 1,700 times, and the more recent GloVe model (Pennington et al., 2014) almost 7,100 times, with the citations per year steadily increasing. With the advent of neural network-based models, most prominently the *word2vec* model (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013), this trend was accelerated even more - in fact, the articles by Mikolov, Chen, et al. (2013) and Mikolov, Sutskever, et al. (2013) have already by far surpassed the original articles on the LSA and HAL models in total citations (currently about 10,000 and 12,000, respectively). Given the success of these models, in terms of performance as well as scientific impact, it is very likely that this trend will explode in the near future.

Furthermore, while working with these models at first required considerable amounts of technical knowledge, applications based on them have become far more accessible over the last years. A critical milestone in this development was the release of the LSA homepage (`http://lsa.colorado.edu/`; see Dennis, 2007), which allows to obtain similarity metrics from LSA models. Additionally, different pieces of software have been released in recent years that allow relatively easy access to vector model-based computations (e.g. Dinu, Pham, & Baroni, 2013; Günther, Dudschig, & Kaup, 2015; Mandera, Keuleers, & Brysbaert, 2017; Pennington et al., 2014; Řehůřek & Sojka, 2010; Shaoul & Westbury, 2010). Given these circumstances, it is therefore critical that these models are properly represented and understood by the scientific community, both from a theoretical and a methodological perspective. As stated on the front page of the LSA homepage: "It is **essential** that you understand the LSA modeling methods before using the applications on this website" (emphasis in original).

However, the need to summarize these models for clear, accessible communication to a general readership, and the rhetorical demand to highlight specific aspects to make arguments or counter-arguments, has at times lead to some rather superficial or incomplete portrayals in the psychological literature. This is especially true from a theoretical perspective. In work not directly focussed on word vector models, they have often been contrasted with allegedly differing theoretical views. For example, in work on the psychological structure of concepts, they are at times introduced as "black-box" models with abstract, non-interpretable dimen-

sional values (e.g., Jones & Mewhort, 2007), while in work on embodied and grounded cognition, they are depicted as language-centric, purely symbolic models of meaning (e.g., Glenberg & Kaschak, 2002; Glenberg & Robertson, 2000).[1] In the context of the specific claims and arguments such studies want to endorse, these references to vector-based models can be appropriate and justified. However, when directed at a general audience that is not necessarily familiar with the large body of specialized literature on these word vector models, this can lead to misrepresentations and snap judgements, which repeatedly crop up in discussions on and assessments of vector space models as cognitive theories, and the research based on them. In the remainder of this article, we will refer to these using the term *misconceptions*. However, we want to stress that by using this term we are not implying that the literature on DSMs is full of errors. We rather refer to portrayals of DSMs that can lead to misconceptions on the side of the readers of such work and in secondary literature.

Apart from that, some claims have been made that were completely accurate at their time and for the models as presented in the original proposals (Landauer & Dumais, 1997; Lund & Burgess, 1996), but have since then been addressed or resolved in extensions or newer models. However, if the initial critical claims appeared in influential articles, they tend to stick in the debate, especially for a general audience that will not necessarily be reached by new technical developments in research explicitly focussed on word vector models. Thus, when earlier claims are taken at face value without considering more recent research, progress in the scientific debate will be hindered.

This phenomenon is further amplified by a rift in the research on word vector models, which is largely split between the fields of machine learning and natural language processing on the one hand (more oriented towards model development, language engineering, and artificial intelligence applications) and cognitive psychology on the other hand (more oriented towards understanding the human cognitive system, predicting human behaviour, and applying measures derived from the models as proxies for cognitive variables), with relatively little communication between the two fields. As a consequence, modelling advancements are sometimes designed as engineering solutions without awareness of theoretical debates in cognitive science. However, insights from cognitive theory can serve as a highly valuable tool for model building, since the human cognitive system often excels in the very tasks that language engineering tries to solve. On the other hand, the theoretical importance of modelling advancements can be overlooked, especially when they are conceived in non-psychological fields. As a consequence, theory building in cognitive science will be hindered if such advancements in model-building are not considered, discussed or interpreted.

Due to these issues, certain arguments concerning vector models of meaning tend to be frequently reported in the scientific discourse, and are discussed and addressed quite regularly. Such repetitions in debates, without a critical assessment of the actual models, their theoretical assumptions, and their more recent advancements, will over time impede scientific progress. We therefore consider it profitable to collect these arguments and to address them thoroughly in the present article. Thus, we aim at providing a state-of-the-art cognitive perspective on these models as theories of human semantic representations. In the present article, we will refrain from delving too deep into the technical details; readers interested in these aspects are referred to the literature cited for the individual arguments. Excellent reviews in this respect are provided by Sahlgren (2006) and Turney and Pantel (2010), as well as by Lenci (2008, 2018), Jones et al. (2015) and Landauer, McNamara, Dennis, and Kintsch (2007).

In the next section, we will first introduce word vector models in more detail, with a focus on their underlying theoretical assumptions, before turning to a collection of common misconceptions about them, which will be discussed thoroughly.

Vector Space Models as Distributional Semantics

So far, we have treated models such as LSA, HAL, Topic Models, GloVe or *word2vec* as a collection of computationally implemented models of semantic representation. However, the commonality of these models goes beyond them all representing meanings as high-dimensional numerical vectors (often referred to as *distributional vectors*, *word vectors*, or *word embeddings*). Actually, all of these models can be seen as specific parametrizations of a unique generalized model, built on the same theoretical foundation: the *distributional hypothesis*, according to which words with similar meanings tend to occur in similar contexts (Firth, 1957; Harris, 1954). Following this hypothesis, there is therefore a correspondence (or, in a more radical view, even an equivalence) between a word's distribution over contexts and its meaning (see also Landauer, 2007). In a strong reading of the distributional hypothesis (described in Lenci, 2008), the contexts in which a word is used not only follow from its meaning, but also determine that meaning.

The power of this hypothesis lies in the fact that applying rigorous operationalizations of "context" allows to quantify a word's distribution and hence, following the distributional hypothesis, its meaning. A very simple operationalization is to define "context" as other words in the same sentence. For example, in the segment *Never gonna give you up, never gonna let you down*, the word *give* occurs two times in the context of *never*, *gonna*, and *you*, and one time in the context of *up*, *down*, and *let*. The word's distribution and therefore its meaning are then represented as the numerical vector (2,

[1] These arguments will be presented in more detail later in the article.

2, 2, 1, 1, 1). If the same contexts are considered for all words, the resulting vectors will all populate a common vector space, the *semantic space*. In realistic applications, where the co-occurrence data is collected from large corpora of natural text, of course more than six different contexts are considered. As a result, the vectors and semantic spaces end up being high-dimensional. Common definitions of contexts are the documents (i.e. sentences, paragraphs or articles) a word occurs in (Landauer & Dumais, 1997; Griffiths et al., 2007) or other words within a fixed-size window around it (Lund & Burgess, 1996; Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013; Pennington et al., 2014).

Hence, the described vector space models all adhere to this specific theoretical approach to meaning. Due to this, they are often referred to as Distributional Semantic Models (*DSMs*), which is a collective term for these models we will be using throughout the remainder of this article. Note that this term carries with it a certain theoretical commitment, in that it implicitly assumes the validity of the distributional hypothesis (otherwise, the models would only be *distributional models*, without being *semantic models*). However, as will be discussed throughout the article, there are a number of arguments in favour of this assumption, which we believe sufficiently justifies the adoption of this widely-used term.

Literature Review

To collect the most common misconceptions about DSMs, we reviewed the description of DSMs in the 1,000 most cited articles in the *Web of Science* database, which themselves cited the original LSA article (Landauer & Dumais, 1997), as of July 2018. These articles were selected as being the most influential and best established scientific works relying on or discussing DSMs. Partly due to the indexing criteria of *Web of Science*, most of these articles belong to the domains of *Psychology (Experimental)* (672) and *Psychology* (231), followed by *Computer Science (Artificial Intelligence)* (404) and *Computer Science (Information Systems)* (192) as well as *Linguistics* (251).

These articles were then reviewed with respect to their description or discussions of DSMs, and findings and results based on these models. We categorized the most common misconceptions into three major topics, each containing more specific sub–issues, that will constitute the main structure of the present article:

1. The implementation of DSMs and the construction of distributional vectors, with a focus on why DSMs are implementations of a cognitive theory rather than statistical tools;

2. Detailed aspects of meaning and information as captured by distributional vectors, with a focus on topics such as interpretability, polysemy, and context-sensitivity;

3. The role of non-linguistic experience in DSMs.

The Implementation of Distributional Semantic Models

We will now turn to the first set of partial descriptions of DSMs in the literature, which are concerned with the exact nature of DSM representations. Alongside this, we describe how traditional, count-based DSMs[2] such as LSA, HAL or GloVe are typically implemented (for comprehensive overviews, see Lenci, 2018; Turney & Pantel, 2010), and what type of information their word vectors actually represent.

Do DSMs conceptualize meaning similarity as word co-occurrences?

In some descriptions, DSMs are presented as (refined) word co-occurrence measures (Barsalou, 1999; Dreyer & Pulvermüller, 2018; French & Labiouse, 2002; Goldberg, Perfetti, & Schneider, 2006; Lau, Goh, & Yap, 2018; McKoon & Ratcliff, 1998; Meteyard, Cuadrado, Bahrami, & Vigliocco, 2012). For example, in very recent studies, Dreyer and Pulvermüller (2018, p. 66) describe DSMs as "distributional learning of word-word correlations from texts", and Kowialiewski and Majerus (2018) explicitly use LSA cosine similarities as a word co-occurrence measure, stating that "LSA measures the extent to which two words co-occur within similar contexts using large corpora" (p. 70). On a more theoretical level, French and Labiouse (2002, p. 316) criticize DSMs because they "take issue with the claim that lexical co-occurrence alone can capture real-world semantics".

It is true that actual word co-occurrences – or, more precisely, word-context co-occurrences - form the data base for DSMs. In traditional DSMs, the first step in the implementation of the model is to construct a word-by-context matrix, either of the word-by-document format (Landauer & Dumais, 1997; Griffiths et al., 2007) or the word-by-word format (Lund & Burgess, 1996). The cell entries of such a matrix are these actual co-occurrence counts. However, DSMs represent word meanings as distributional vectors, and therefore complete rows of such a matrix, and not single cell entries. As a consequence, two words are similar in meaning not because of their mutual co-occurrence score (which would correspond to such a single cell entry, and only for word-by-word matrices), but rather if they have similar global distributional *patterns* over all contexts. Due to this, synonyms, which tend to very rarely co-occur directly, will have very similar meanings in DSMs (Sahlgren, 2008). This property of DSMs also

[2]Recent prediction-based models such as *word2vec* are implemented using a different architecture, but this architecture in many respects is based on a similar rationale as the traditional models (Levy & Goldberg, 2014b). We will describe the *word2vec* algorithm in more detail later in the article.

	patient	*hospital*	*medicine*	*physician*	*doctor*
physician	4	5	3	0	0
doctor	10	10	7	0	0

$$\cos(physician, doctor) = .995$$

Figure 1. Simplified word-by-word co-occurrence matrix. As can be seen, the actual co-occurrences of the words *physician* and *doctor* are extremely low (zero); however, their distributional patterns – their respective rows in the matrix – are very similar.

gives rise to the important notion that words that never occur together can nevertheless end up with very similar meaning representations (Landauer & Dumais, 1997), as can be seen in Figure 1.

Do DSMs describe word meanings as co-occurrence patterns?

Other descriptions, while recognizing that DSMs capture global instead of local co-occurrences, still present them as co-occurrence models (Borghi, Glenberg, & Kaschak, 2004; ?, ?; Rommers, Dijkstra, & Bastiaansen, 2013; Van Dam, Rueschemeyer, & Bekkering, 2010; Van Herten, Chwilla, & Kolk, 2006). This is typically expressed along the lines that DSMs give a "semantic similarity measure that computes how often two words co-occur with the same set of other words" (Rommers et al., 2013, p. 766), that "LSA values do not reflect how often a pair of words co-occur, but rather how often they co-occur with the same words" (Van Tiel, Van Miltenburg, Zevakhina, & Geurts, 2016, p. 159) or that they are an "objective measure of co-occurrence in related contexts" (Borghi et al., 2004, p. 867). In the context of these articles, such descriptions are clearly intended as a short but comprehensible summary of DSMs. And, indeed, such descriptions would be completely accurate if the word-by-context matrix was the final step in the implementation. However, in actual implementations, the raw co-occurrence vectors are typically subject to further processing.

Typically, weighting schemes are applied to the raw co-occurrence counts, such as positional weighting depending on the number of intervening words (in word-by-word models such as Lund & Burgess, 1996), log-entropy weighting (Martin & Berry, 2007) or Pointwise Mutual Information (PMI, Church & Hanks, 1990). For example, PMI is defined as

$$PMI = \frac{P(a,b)}{P(a) \cdot P(b)}$$

with $P(a,b)$ being estimated via the co-occurrence frequency and $P(a)$ and $P(b)$ via the global word frequency (or number of words in a document, if documents serve as context). On the one hand, the purpose of such weightings is to adjust for word frequency effects – obviously, very frequent words would result in high co-occurrence values, which then would heavily influence the direction of the word vectors. On the other hand, this also has the effect that the basis for the word vectors is not the raw contiguity between a word and a context, but rather the informativity of the relation between them – their contingency. For example, the raw co-occurrence frequency between *bark* and *dog* is lower than the co-occurrence frequency between *dog* and *the*, because of the high frequency of *the*. The former pair, however, will have a higher PMI, clearly capturing the significant degree of informativity between its elements.. And this is in fact highly desirable for a theory of semantic memory, as it is in line with general learning theories emphasizing that contingency, not contiguity, drives the learning of associations between stimuli (see, for example Murdock, 1982; Rescorla & Wagner, 1972).

These weighted co-occurrence vectors can then be forwarded to a further processing step: dimensionality reduction (for earlier applications of dimensionality reduction techniques in research on semantic representations, see for example Heider & Olivier, 1972; Osgood, 1952). The core principle of such techniques is to identify redundancies and mutual constraints in high-dimensional data patterns in order to extract underlying factors. Then, factors that account only for little variability are dropped, so that a large part of the variability in the data can be explained with lower-dimensional representations. As a very rough example, large parts of the three-dimensional object "sheet of paper" can be described using a two-dimensional representation. Typical dimensionality reduction techniques for DSMs are Singular Value Decomposition (SVD; Martin & Berry, 2007), Latent Dirichlet Allocation (LDA; Blei, Ng, & Jordan, 2003) or Non-negative Matrix Factorization (NMF; Arora, Ge, & Moitra, 2012).

Dimensionality reduction has been crucial to the success of DSMs, and is associated with large gains in performance (Landauer & Dumais, 1997; Bullinaria & Levy, 2012, but see Recchia & Jones, 2009). It is therefore to be considered a core feature of DSMs, also from a theoretical point of view (in fact, this is the reason for the expression *Latent* Semantic Analysis): Word meanings are not obtained by just observing word-context co-occurrences, but by abstracting and generalizing from them and by establishing higher-order representations relying on global information. As stated by Landauer and Dumais (1997, p. 217), the "relation between any two representations depends not only on direct experience with them, but with everything else ever experienced". This transition from first- to higher-order relations is claimed to reflect the transition from episodic memory, capturing concrete instances of co-occurrence of entities, to semantic memory, capturing more fundamental, conceptual relations between them (see Landauer & Dumais, 1997).

Are DSMs just language statistics?

As introduced previously, DSM algorithms are statistical procedures applied to large collections of text in order to obtain quantitative representations of word meaning. Therefore, DSMs are at times described as language statistics, "a computer program that computes an index of the relatedness between sets of words on the basis of occurrences in similar contexts" (Kaschak & Glenberg, 2000, p. 521), or reduced to being methodological tools (Perfetti, 1998).

However, the seminal work on LSA by Landauer and Dumais (1997) already puts forward strong arguments in favour of DSMs as explanatory theories rather than methodological tools. In fact, DSMs are set up as a theory explaining how semantic representations are acquired: These representations can be extracted from a given distributed input, and from contingencies in the experience of a learner/speaker. DSMs start from the theoretical assumption that word meanings are inferred from the contexts in which the words occur (based on the distributional hypothesis). In essence, this is a specification of the assumption that word meanings are acquired through experience, as stated by Jenkins (1954, p. 112): "intraverbal connections arise in the same manner in which any skill sequence arises, through repetition, contiguity, differential reinforcement".

In a more general perspective, DSMs thus stand in the tradition of learning theories postulating that humans are excellent in capturing statistical regularities in their environments (Anderson & Schooler, 1991) and extracting information from them (see also Günther, Smolka, & Marelli, 2018). This ability can be observed in many different domains, such as visual pattern recognition (Kirkham, Slemmer, & Johnson, 2002), procedural knowledge (Lewicki, Czyzewska, & Hoffman, 1987), or social cognition (Lewicki, 1986), but has also been found to play a vital role in language acquisition (Saffran, Aslin, & Newport, 1996; Saffran, 2003).

Starting from this assumption, a model is built that includes cognitively motivated processing steps (see the previous paragraph). In contrast to many verbal theories proposed in cognitive science, DSMs postulate a computational implementation of the underlying theory, and this implementation includes statistical processing steps. However, the fact that DSMs come with an implementation based on actual linguistic data should not be considered a flaw. On the contrary, it is a major advantage of these models, also from a theoretical point of view: This implementation allows researchers to derive quantitative hypothesis so that the theory's predictions can actually be subjected to rigorous empirical tests, which is in turn crucial for model adjudication.

Are DSMs psychologically implausible learning models?

When getting familiar with the traditional count-based algorithm implementing DSMs such as LSA and HAL, as described above, a very common reaction is to point out that these models make very questionable assumptions about the learning processes leading to semantic representations. These shortcomings of traditional DSMs have already been recognized and acknowledged (e.g. Sloutsky, Yim, Yao, & Dennis, 2017), also by proponents of DSMs (Lemaire & Denhière, 2004). Traditional DSMs are based on batch-learning algorithms, where a matrix storing all information is processed "at once", using computationally demanding techniques. The results are static word meanings estimated from a fixed corpus of experience. And the assumption that humans collect large amounts of co-occurrence data before at some point turning them into semantic representations is not very convincing (Hoffman, McClelland, & Lambon Ralph, 2018). In reality, word meanings are not static, but can change dynamically and incrementally with new experience (twenty years ago, our semantic representations for *cloud* or *phone* were probably very different from now). Therefore, it is usually claimed that models such as LSA are "not a model of learning" and have "no obvious way of accounting for developmental changes in word learning" (Sloutsky et al., 2017, p. 6). However, when only LSA is presented as a prototypical DSM and as the basis for such an argument, without considering more recent developments, this gives the incorrect impression that DSMs in general cannot deal with incrementality.

In principle, a traditional DSM algorithm such as LSA can cope with incrementality by updating the co-occurrence matrix with each new input and re-running the algorithm. However, this solution is quite unsatisfactory (see Recchia & Jones, 2009): Firstly, the algorithm is quite computationally demanding, especially because of the dimensionality reduction step, and hence would put an enormous strain on cognitive resources during language processing. Even more worrying, such a mechanism would require a language user to have the whole raw co-occurrence data, as stored in the initial matrix, available at any time, basically reducing the whole point of dimensionality reduction to absurdity.

However, as this criticism has been brought up early within the DSM community, it has also been addressed there. Martin and Berry (2007) described an algorithm for "folding in" new documents into an existing LSA vector space. However, this still requires a static semantic space to be constructed at some point in time. Addressing the issue of incrementality more directly, new DSM algorithms have been proposed that explicitly implement incremental learning of semantic representations. These models do not start from co-occurrence matrices, but from random vectors whose elements are then updated with each new encountered text unit (Jones & Mewhort, 2007; Sahlgren, 2005), also by using learning algorithms derived from psychological learning models (for example the paired-associate learning mechanism by Murdock, 1982, in the BEAGLE model,

Jones & Mewhort, 2007).

More recently, prediction-based DSMs have been introduced that estimate word vectors using a neural network architecture with one hidden layer (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013). In this architecture (see Figure 2 for a graphical description), words serve as the input layer from which its context words – the output layer – are predicted (or vice versa). The activation values of the hidden layer for a given word in the input layer is then taken as its associated semantic vector. This network is trained on a corpus, word by word, and so the incremental development of distributional vectors over time is an inherent property of this algorithm. In the psychological literature, it has been demonstrated that these predicition-based models are mathematically equivalent to psychologically plausible learning models (Hollis, 2017; Mandera et al., 2017), such as the Rescorla-Wagner model of reinforcement learning (Rescorla & Wagner, 1972; see also Gluck & Bower, 1988; Sutton & Barto, 1981).

It has also been shown that the *word2vec* model outperforms traditional count-based models in a variety of tasks, including the prediction of human behaviour (Baroni, Dinu, & Kruszewski, 2014; Mandera et al., 2017; Pereira, Gershman, Ritter, & Botvinick, 2016). Furthermore, a recent study by Lazaridou, Marelli, and Baroni (2017) has shown that such models can learn the meaning of novel words from linguistic context in a human-like fashion, even from very limited exposure to these words. This nicely highlights the advantages of integrating technical advancements in model building with insights from cognitive theory, and the mutual benefit to both fields when they go hand in hand.

Moreover, while such incremental and prediction-based models may seem to be far off from the traditional DSM algorithm described above, Levy and Goldberg (2014b) showed that they can be described as performing an implicit matrix factorization of a PMI-weighted word-context matrix. This suggests that even traditional DSM algorithms might not be as implausible as they appear to be at first glance. This is mirrored in the claim already put forward by Landauer and Dumais (1997) that the matrix-factorization algorithm describes *what* the system does (the *computational* level of description; Mandera et al., 2017) rather than giving an accurate model of *how* this is achieved (the *algorithmic* level of description, incorporated more thoroughly in prediciton-based models).

Are DSMs language-engineering tools rather than psychological models?

DSMs have two major fields of application: While there is a large body of research in cognitive science applying and investigating these models, they are even more widely applied in computer science and computational linguistics. The existence of a vast body of literature in the fields of machine learning, natural language processing and artificial intelligence research focused on solving language-related tasks suggests, on an implicit level, the view that DSMs are an engineering tool rather than a cognitive theory: In this engineering-oriented view, there is typically little emphasis on the underlying theory and the cognitive plausibility of DSMs as a model of human semantic representations, as these aspects are not necessarily required to build well-performing models and algorithms for these given tasks.

While DSMs might be valuable in order to engineer word meanings, this does not automatically qualify them as plausible psychological models (Glenberg & Mehta, 2008; Perfetti, 1998), and "the utility of vector-space models for understanding human semantic abilities remains in question" (Rogers & Wolmetz, 2016, p. 124–125). Sahlgren (2006, p. 134-135) explicitly states that it "cannot be stressed enough that the word-space model is a *computational* model of meaning, and *not* a psychologically realistic model of human semantic processing" (emphasis in original) and that they represent "not the meanings that are in our heads, and not the meanings that are out there in the world, but the meanings that *are in the text*" (Sahlgren, 2008, p. 46, emphasis in original). Hence, they might prove highly useful for machine applications and artificial cognitive systems, but not as a psychological model for human semantic representations.

However, these objections can be met with both theoretical and empirical arguments. The theoretical arguments concerning how DSMs are conceived but also implemented as psychological models were already outlined in the previous paragraphs. They are based on a cognitive hypothesis of meaning (Lenci, 2008), and make cognitive assumptions about how these meanings are acquired (Hollis, 2017; Landauer & Dumais, 1997; Mandera et al., 2017) – through repeated experience (Jenkins, 1954). The nature of a theory being a "cognitive" one is first of all determined by its *scope*, rather than its validity.

However, DSMs could still be bad cognitive theories; whether the hypotheses and assumptions of DSMs as a cognitive theory of semantic representation are valid, and therefore whether they are good or useful theories, is subject to empirical investigation. On the empirical side though, there is a large collection of results showing that DSMs can account for and predict human behaviour in a wide range of semantic memory-related tasks (Baroni, Dinu, & Kruszewski, 2014; Pereira et al., 2016; see also Landauer et al., 2007; Lenci, 2008). These include (among others, see the studies cited throughout the present article) categorization tasks (Baroni & Lenci, 2010; Louwerse, 2011), synonym tests (Bullinaria & Levy, 2012; Landauer & Dumais, 1997), similarity judgements (Bruni et al., 2014), free associ-

"... the *red player* ***defeats*** *blue opponents"* ...

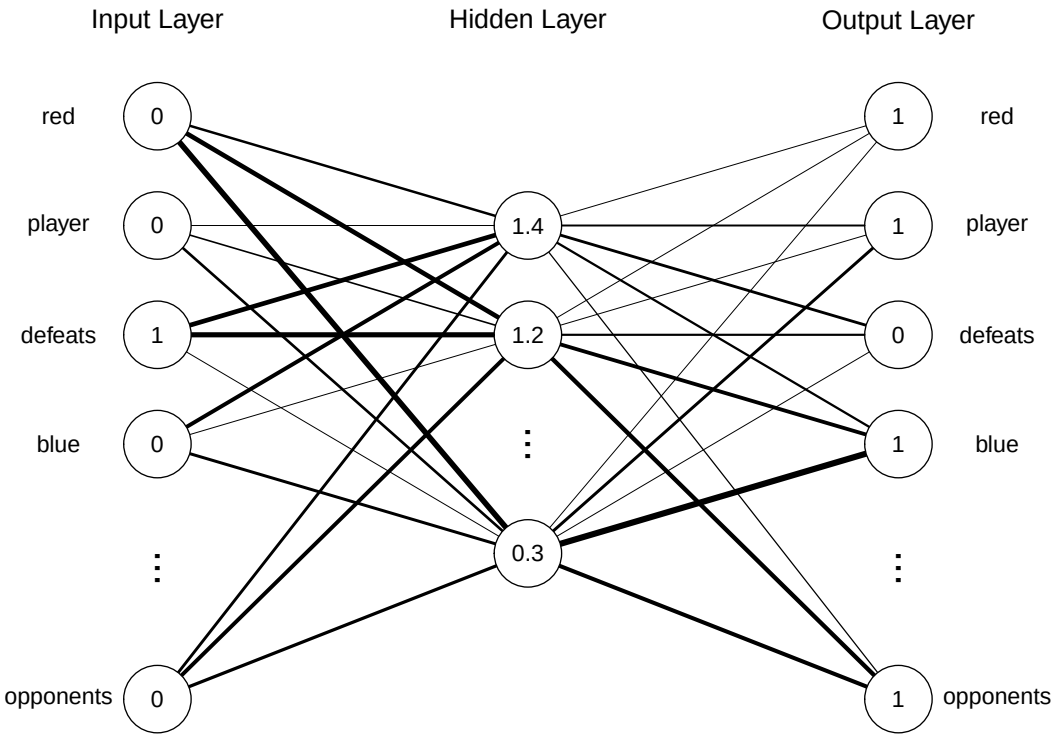

Figure 2. A snapshot of the *word2vec* model architecture, in the training step of processing the target word *defeats* in the utterance "... the *red player beats blue opponents"* ... (with a window size of 2). The widths of the edges represent the current weights in the network, which are updated in every training step. The activation of the hidden layer (1.4, 1.2, ..., 0.3) is the word vector representing *defeats* in the current state of the model, and will change as the weights between input layer and hidden layer are updated. In the skip-gram version of the model (depicted here), the context words are predicted from the target word. In the cbow version, the target word is predicted from the input instead (i.e., the input and the output layer are switched).

ation tasks (Nematzadeh, Meylan, & Griffiths, 2017), semantic priming (Günther, Dudschig, & Kaup, 2016; Jones et al., 2006; Mandera et al., 2017), concept acquisition (Ouyang, Boroditsky, & Frank, 2017; Lazaridou et al., 2017) and text comprehension (Landauer, Foltz, & Laham, 1998). Of course, a high performance in predicting human behaviour qualifies a model as a good description at the computational level, and not necessarily on the algorithmic level of *how* this performance is achieved (Mandera et al., 2017). However, even in this case, DSMs already fulfil the criteria to be considered as psychological models of the nature of semantic representations and the structure of semantic memory, similar to other models of semantic memory (see Jones et al., 2015).

Interestingly, more recent work in the DSM literature has more thoroughly considered the question of acquisition. DSMs such as BEAGLE and *word2vec* have been shown to incorporate psychologically plausible learning mechanisms (Murdock, 1982; Rescorla & Wagner, 1972) to create their word representations, raising their algorithmic plausibility in comparison to earlier models (Hollis, 2017; Jones et al., 2015; Mandera et al., 2017, see the previous section).[3]

Following from these arguments, DSMs should not be refuted from the outset as being mere engineering tools, as they are formulated as cognitive theories, rely on psychologically plausible assumptions, and are able to account for behavioural data. However, we want to emphasize again that

[3]We should also note at this point that, when the issue of cognitive plausibility at the algorithmic level is taken seriously, the source corpus from which the distributional vectors are estimated plays a more important role than in the context of engineering solutions. For a cognitive model, the corpus should itself be plausible in the sense that it serves as a good reflection of the actual experience made by a human, preferrably in structure as well as in size.

they are qualified as psychological theories by their scope, and not necessarily by their quality. In this section we have focussed on empirical phenomena that DSMs can account for; yet, other psychological theories of semantic representation have been argued to account for phenomena that DSMs cannot explain (see, for example, Glenberg & Robertson, 2000). Thus, we argue that DSMs are serious contenders as psychological theories of semantic representation, but their quality as such – as for any other theory – is subject to scientific evaluation.

Aspects of Meaning

In the previous section, we have described the algorithms used to create semantic spaces. In this section, we will now turn to common misconceptions related to the implications of the vector representation format for word meanings.

Are distributional vector dimensions uninterpretable?

As a consequence of the application of dimensionality reduction in the model architectures described above, the dimensions of distributional vectors represent abstract, or latent, semantic dimensions (Landauer & Dumais, 1997). Jones and Mewhort (2007, p. 2) summarize this well (referring to the LSA model) by stating: "In a semantic space model, the features that represent a word are abstract values that have no identifiable meaning in isolation from the other features. Although a particular feature of bird in a feature list might be "has wings," the presence of which has birdlike meaning on its own, the meaning of bird in a semantic space model is the aggregate distributed pattern of all the abstract dimensions, none of which has interpretable meaning on its own." Hence, distributional vector dimensions are usually described as not meaningful or interpretable (e.g., Borghesani & Piazza, 2017; Hansson, Bååth, Löhndorf, Sahlén, & Sikström, 2016), rendering distributional vectors "devoid of content" (Rogers & Wolmetz, 2016, p. 124).

This is an especially troublesome property for many linguistic theories of semantics, which often heavily rely on well-defined semantic features (see Bierwisch, 2011; Nida, 1979; Johnson, 2008), but many psychological theories of word meanings and concepts are also founded on interpretable features (e.g. Collins & Loftus, 1975; McRae, Cree, Seidenberg, & McNorgan, 2005; Smith & Medin, 1981).[4]

For many DSMs, such as LSA, HAL, or *word2vec*, these claims are completely accurate. However, we also want to point out noteworthy exceptions and developments. A quite prominent case of DSMs with interpretable dimensions are Topic Models (Griffiths et al., 2007). At the very foundation of these models lies the core theoretical assumption that texts/documents are generated to cover certain *topics*. In Topic Models, Latent Dirichlet Allocation (Blei et al., 2003), a dimensionality reduction technique, is applied to a word-by-document matrix to detect these topics. As a result, each word vector represents a probability distribution over semantic topics (models using Non-Negative Matrix Factorization as a dimensionality reduction technique are based on the same idea; Dinu & Lapata, 2010; Lee & Seung, 1999). Griffiths et al. (2007) provide examples of these topics by presenting the words that have the highest probability scores on different topics: For example, the highest-value words on one topic were *play, stage, audience, theater,* and *actors*, on another topic they were *hypothesis, experiment, scientific, observations,* and *scientists*, and on another topic they were *class, Marx, economic, capitalist,* and *socialist*. These topics can be labelled (similar to labelling latent factors after a factorial analysis) – in this case using labels such as *drama*, *science*, and *Marxism* – making each dimension of the semantic space interpretable (but see Chang, Gerrish, Wang, Boyd-Graber, & Blei, 2009, for criticism concerning the objectivity of such post-hoc topic labels). In a similar fashion, other studies were able to derive interpretable dimensions from traditional LSA spaces by applying mathematical procedures such as varimax rotation (Evangelopoulos, Zhang, & Prybutok, 2012) or base transformations (Olmos, Jorge-Botana, León, & Escudero, 2014) to these spaces .

Other studies have shown that interpretable information can be extracted from initially uninterpretable distributional vectors. Hollis and Westbury (2016) applied a principal component analysis to a semantic space obtained through *word2vec*, and found that the principal components obtained can be identified with basic-level semantic dimensions such as concreteness, or the affective dimensions of valence, arousal, and dominance already postulated by Osgood, Suci, and Tannenbaum (1957). In general, understanding what kind of semantic information is captured by distributional vectors is a topic of intense debate and current research. For instance, Tsvetkov, Faruqui, Ling, Lample, and Dyer (2015) show that semantic type information (such as ANIMAL or MOTION) can be, at least partially, recovered from distributional vector dimensions. Further, Durda, Buchanan, and Caron (2009) demonstrate that feature norms produced by human speakers can be predicted from distributional vectors, using a neural network mapping approach. And in a very recent study, Sommerauer and Fokkens (2018) show that distributional vectors capture semantic properties related to the functionality of the denoted concepts, by employing a supervised classification algorithm that estimates from the vector dimension values whether or not a given word has a given semantic property, such as *is dangerous*. These studies shows that distributional vectors in principle

[4]Note however that the reliance on well-defined features can also be framed as a problem of such theoretical approaches rather than a problem of feature-free models of meaning; see for example Westbury (2016) for arguments along this line.

encode semantically interpretable information, which can be extracted by using adequate methods.

Does each distributional vector represent a single, fixed symbolic meaning?

In a semantic space, each unique word form is associated with exactly one high-dimensional numerical vector representing its meaning. In reality, however, a very large proportion of words have multiple meanings. For example, some words are homonyms: A *mouse* can be an animal or a piece of computer hardware. Beyond this, the issue of polysemy, where words have multiple related senses, is even more widespread: A *paper* can be a physical piece of paper, but also a text printed on this paper, and even a piece of text without being printed on paper. And the fact that meanings are context-sensitive is virtually ubiquitous: For the exact same object described by the word *newspaper*, the aspect of carrying printed text is relevant in the sentence *I read the newspaper*, even if it is an online newspaper. On the other hand, in the sentence *I use the newspaper to protect my face from the rain*, the aspect of being a solid physical object is relevant (Glenberg & Robertson, 2000).

Since DSMs represent word meanings as a single numerical vector with fixed values, Glenberg and Robertson (2000) argue that they cannot appropriately capture this context-sensitivity. A similar argument is made by French and Labiouse (2002), who maintain that distributional vector representations are not context-sensitive because they are only an average representation of the word's distribution. Accordingly, they conclude that DSMs cannot accurately represent metaphorical meanings, such as *wolf* in the sentence *John is a real wolf with the ladies* – which does not imply that he turns into an actual canine in female company. Thus, it has been concluded that multiple meanings, senses or interpretations cannot be represented via a one single vector.

However, this argument neglects the fact that distributional vectors are distributed representations that encode a word's learning history, rather than being a symbolic meaning representation (see Westbury, 2016). Thus, if a word has multiple meanings or senses, these will be encoded in the vector and its position in the semantic space; in that, the observation by French and Labiouse (2002) that vectors are average representations is correct. However, this does not imply the conclusion that multiple meanings or senses can no longer be retrieved from these vectors. Figure 3, created using the R package *LSAfun* (Günther et al., 2015), provides a intuitive basis for this argument using the *mouse* example (Camacho-Collados & Pilehvar, 2018), and demonstrates that semantic spaces are structured in a way that allows to retrieve different meanings and senses from a single word vector, by considering its position and the relative positions of other words in the semantic space (for more detailed approaches, see Camacho-Collados & Pilehvar, 2018; Heylen, Wielfaert, Speelman, & Geeraerts, 2015). In a different example, Griffiths et al. (2007) show that different semantic dimensions of their distributional vectors capture different contexts in which words are used – and by extension, different meanings and senses (see Lee & Seung, 1999, for a similar demonstration). Hence, DSMs seem to well account for different subordinate meanings as a function of contextual information.

Additionally, various computational methods have been shown to consistently outperform statistical baselines in disambiguating different word senses and meanings from distributional vectors (see Camacho-Collados & Pilehvar, 2018), for example by relying on their respective similarities to multiple different pre-determined word categories, or clusters (Boleda, Padó, & Utt, 2012; Pantel & Lin, 2002). Although the introduction of such pre-determined categories is potentially problematic from a psychological perspective, as actual speakers have to learn these categories from experience in a bottom-up and dynamic way, rather than obtaining them top-down from an outside source, such studies still demonstrate that distributional vectors encode information that can be harvested to disambiguate different senses.

Pursuing a different line of argumentation, Westbury (2016) argues that, even for words traditionally considered monosemous, the assumption of single symbolic word meanings can be called into question. It is not unusual that two entities that share only very few features – or sometimes even no features at all – are assigned the same label: Two patients diagnosed with *depression* can have a disjunct set of symptoms, or several experiences one describes as *happy* can have nothing in common. Due to this, Westbury (2016) argues that word meanings should be seen as mappings from particular experiences and instances onto a common label rather than well-defined feature definitions – which fits perfectly with distributional vector representations of word meaning.

Concerning the issue of context-sensitivity of word meanings and senses in sentences, Jones and Mewhort (2007) point out that multiple meanings of a word can be stored within a single vector representation – as shown above – and that different meanings can emerge from a single representation depending on the word's context (as postulated in different psychological theories of discourse comprehension; Kintsch, 1988; Tabossi, 1988). As a direct demonstration, Kintsch (2001) proposed a model based on his construction-integration theory (Kintsch, 1988) that computes a vector representation for predicate-argument constructions (such as

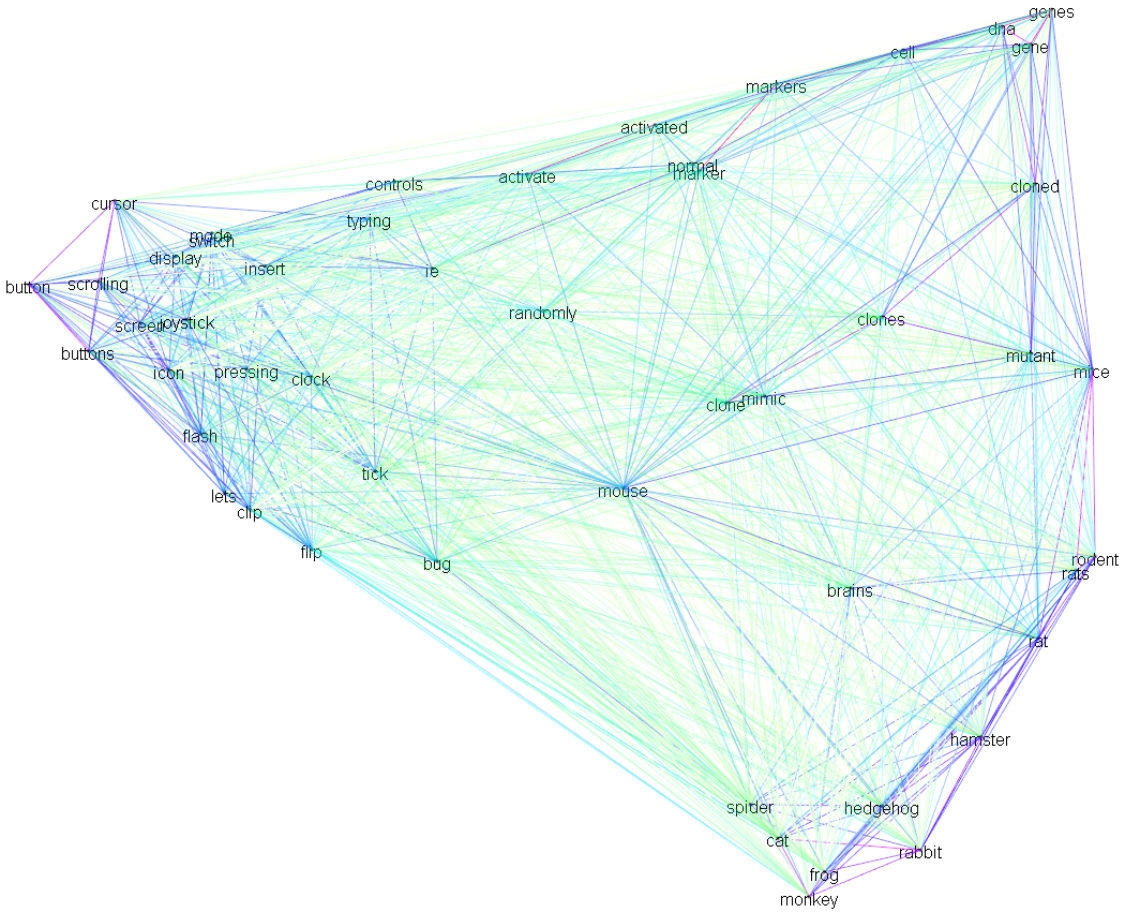

Figure 3. A two-dimensional projection of the 50 nearest neighbours of mouse in an actual semantic space. Semantic similarity is represented by the distance between the words as well as the colour of lines connecting them. As can be seen, the neighbourhood is clustered into different meanings and senses (computer mouse, mouse as an animal, mouse as a laboratory animal), with high within-cluster similarities but low between-cluster similarities.

SHARK(LAWYER) in *My lawyer is a shark*). This model updates the predicate vector by adding vectors that are close to *both* the predicate and the argument to it, and is therefore argued to capture the meaning aspects of *shark* relevant in this context (being vicious, rather than being a fish). Kintsch (2000) also argues that this model can offer an account for metaphor comprehension, and Jorge-Botana, León, Olmos, and Hassan-Montero (2010) use this model to extract different senses from polysemous words. In a similar spirit, Baroni and colleagues showed that different senses of adjectives are activated depending on the noun they are used with (using compositional methods described in more detail later in the article; Baroni, Bernardi, & Zamparelli, 2014; Baroni & Zamparelli, 2010): An *old tomato* is a *rotten tomato*, *old ruins* are *ancient ruins*, and an *old belief* is a *traditional belief*.

At this point, it might be objected that the context-sensitive modification of one vector is an extension to original DSM representations. However, it should be noted that only standard distributional vectors are employed in these approaches, and it can therefore only work because these vectors already encode all necessary information. In fact, the possibility of integrating LSA with context-models such as the construction-integration model was already pointed out and discussed in the original article by Landauer and Dumais (1997).

Do DSMs only represent isolated word meanings without relational structure?

In DSMs as outlined so far, word vectors are constructed on the basis of surface-level word co-occurrences, without consideration for the relations or syntactic dependencies between the words (as discussed by Hansson et al., 2016). Although moving window models such as HAL or *word2vec*

encode some syntax, this takes place on a very rudimentary level. However, languages typically use syntactic dependencies to express the relations between words, which are often the focus of logic and ontological theories of meaning, which in turn are central to linguistic analyses of semantics (see Annesi, Croce, & Basili, 2013). In addition, DSMs also seem to largely ignore relations between words in their meaning representations: Since DSMs represent word meanings as vectors in a semantic space, research has focussed on simple similarity metrics such as cosine similarity. Thus, DSMs appear to represent word meanings as isolated elements, without actually considering the relations between them, neither in encoding nor in representation. Thus, they appear to miss out on essential aspects of semantics.

However, both issues of encoding and representation have been addressed within the DSM framework, in different lines of research. Using the BEAGLE model which is explicitly designed to encode word order information during learning, Jones and Mewhort (2007) demonstrate that the nearest "left" neighbors for words (*luther*, *barbaric*, and *burger* for *king*) differ from their nearest "right" neighbors (*midas*, *lear*, and *henry* for *king*). Thus, order information can be retrieved from distributional vectors.

Another line of research has directly focussed on building DSMs that explicitly consider syntactic dependencies in the input data and thus already in model building (e.g. Biemann & Riedl, 2013; Lin, 1998; Padó & Lapata, 2007). These models rely on dependency-parsed corpora, and in essence consider words as co-occurring in the same context if there is a specific syntactic relation between them. For example, in the sentence *man bites dog*, *dog* occurs one time as the object of *bite*, but in *dog bites man*, *man* occurs one time as the object of *bite*. Padó and Lapata (2007) demonstrated that DSMs built from syntactic dependencies can predict behavioural data such as semantic priming, detect synonyms, and disambiguate different word senses (showing again that distributional vectors don't represent a fixed symbolic meaning). Another DSM explicitly designed to capture properties of and relations between concepts, such as that a *tiger* is *in* a *jungle* or *has stripes*, was proposed by Baroni, Murphy, Barbu, and Poesio (2010). This model infers these properties by employing a part-of-speech tagged corpus and considering explicitly the type of linguistic construction in which pairs of words co-occur.

However, in this context, Landauer (2007) takes a strong position in arguing that the importance of syntactic dependencies and grammatical relations for mental representations of word meanings is only marginal when compared to the co-occurrence of words, and argues in favour of bag-of-word models[5] of meaning such as LSA. Indeed, recent research provides a proof of principle that also standard DSM vectors can encode structured information, which can be uncovered by using more refined, asymmetrical similarity measures instead of the typical symmetrical cosine information (Kintsch, 2014; Lenci & Benotto, 2012). Kintsch (2014) developed an asymmetrical similarity measure for which, for example, the similarity between *China* and *Korea* is rather low, while the similarity between *Korea* and *China* is very high, reflecting behavioural patterns typically observed in human free associations. In other studies more oriented towards linguistic relations between words, it has been shown that relations such as hypernymy and hyponymy (i.e. set-subset relations: every *dog* is a *mammal*, but the opposite is not true) can be inferred from distributional vectors by using asymmetrical similarity measures (Lenci & Benotto, 2012).

However, semantic relations between words clearly go beyond the quite coarse relations outlined here: For example, the pairs *skin* and *body* as well as *bark* and *tree* both share a COVERS relation. This issue of automatically identifying an open-ended class of semantic relations has been addressed in studies on analogies. In a famous example, Mikolov, Yih, and Zweig (2013) found that the nearest vector to *king - man + woman* is *queen*, suggesting that *king* and *man* share the same relation as *queen* and *woman*. Apart from this illustratory example, the possibility to recover open-ended relational similarities between words from their distributional vectors was systematically confirmed in several studies (Mikolov, Yih, & Zweig, 2013; Turney, 2006; Zhila, Yih, Meek, Zweig, & Mikolov, 2013, see Levy & Goldberg, 2014a).

Do DSMs only capture linguistic knowledge about word meanings?

As described earlier, the data from which DSMs are built typically consists of large amounts of natural text. The general objective of such algorithms has been described by Buitelaar, Cimiano, and Magnini (2005, p.4) as "the acquisition of explicit knowledge implicitly contained in (textual) data". However, in both linguistics and cognitive science, researchers often differentiate between linguistic knowledge on the one hand (what would be considered the semantics of a word), and non-linguistic knowledge on the other hand, derived from our knowledge about the world (Jackendoff, 2003; Lang & Maienborn, 2011; although this differentiation is also debated, see Hagoort, Hald, Bastiaansen, & Petersson, 2004). For example, the fact that the US president is married is part of our world knowledge, while the fact that a husband is married is part of the semantics of the word. Similarly, the fact that being a woman typically is a necessary

[5] LSA is described as a bag-of-word model since it takes as input only a matrix specifying the word-by-document occurrences, without any concern of their structure. It would derive exactly the same distributional vectors if the order of words within a document was randomized or sorted alphabetically.

condition to be a mother is part of our semantic knowledge, while the information that being woman raises the likelihood of being under-privileged in many human societies is part of our world knowledge. Since traditionally DSMs are built from language data, it could be assumed that they only capture linguistic knowledge. Following this distinction, DSMs have been used in empirical studies to exclusively account for linguistic, semantic knowledge in research on linguistic vs. extra-linguistic knowledge (Dudschig, Maienborn, & Kaup, 2016).

However, such a conclusion would underestimate the knowledge about word meanings that is implicitly encoded in natural text, as speakers often use language to communicate about their view on the world, and to express their concepts, assumptions, norms and ideas. Accordingly, Caliskan, Bryson, and Narayanan (2017) have shown that DSMs can in fact draw implicit information related to social and cultural biases from text data: They used DSM vectors to compute word similarities of items used in the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998), in order to approximate reaction times (with higher cosines corresponding to faster reactions). These measures showed the same biases as humans in several domains, such as gender biases with respect to career choices and societal roles (see also Lenton et al., 2009, for an earlier DSM-based approach to this issue), attitudes towards race, or biases towards physical vs. mental health issues. These results were further corroborated in a study by Bhatia (2017b).

In another recent study on world knowledge-based judgements, Bhatia (2017a) employed DSMs similarities to measure associations between mental representations. With this approach, Bhatia demonstrated that DSMs can predict human performance (and biases) in a wide range of high-level judgement tasks, including the conjunction fallacy (Tversky & Kahneman, 1983), base rate neglect (Kahneman & Tversky, 1973) and the recognition heuristic (Gigerenzer & Goldstein, 1996).

Such studies show that the information captured by DSMs goes beyond purely linguistic semantic knowledge, into the domain of (implicit) world knowledge and cultural specificities.

Can DSMs account for productivity of meaning?

Another feature of word meanings in natural language settings is that they can be very productive (depending on the specific language). For example, meanings such as *objectify* can be created from *object* via affixation, and words such as *air* and *port* can be combined into compounds (*airport*). Most importantly, however, word meanings can be created and words that have never been used or experienced before can be readily understood by a language user. This is particularly true for the above-mentioned cases of complex expressions: It is quite easy to understand what it means to *twitterify* politics, or what a *like-addiction* is supposed to be. And language users have no problem in using existing words in novel ways, as in the *newspaper* example in the previous sections.

However, the DSM algorithm runs on a given corpus of text. If a word is not included in such a corpus, no vector representation will be derived. And if a word is never used in a specific sense (i.e. if it is not used in specific contexts), this will in turn affect the vector representation itself. In both cases, incremental addition of new experience as described earlier is not going to help, since the argument is that humans can deal with meanings that they have never experienced or used before. Glenberg and Robertson (2000, p. 397) therefore notice that DSMs cannot deal with novel uses of known meanings, but that a "computational model should be able to account for material beyond its training set. It is especially important that a theory of language and meaning be flexible and productive beyond its training set because humans are flexible and productive." The problem of the non-productivity of the traditional DSM algorithm has also been recognized elsewhere by other authors (Lynott & Ramscar, 2001; J. Mitchell & Lapata, 2010).

However, the issue of novel uses of existing words can be addressed within the DSM framework. As argued in a previous section, several studies have shown that distributional vectors can encode various aspects of word meaning, including semantic features (Durda et al., 2009) and functionality-related semantic properties (Sommerauer & Fokkens, 2018). This does not require that all this information is explicitly mentioned in the text; it can be inferred from the distributional similarity to other words denoting concepts that share the same properties and features. As an example, if you learn that you can build boxes out of something called *polystyrene* to transport pasta and curries, you can easily infer that you could also build something like an umbrella out of this material. This works even if the possibility was never explicitly mentioned, since the distributional use of *polystyrene* is similar to that of other water-proof material such as *plastic* or *polyethylene*. Depending on the context in which a word is used, this relevant dimension can then be emphasized through context-sensitive updating (see Kintsch, 1988, 2001, as discussed in a previous section).

This still leaves the issue of the meaning of novel words. How can DSMs represent meanings for words that are not part of a corpus? The answer can be found in the recent emergence of *compositional* DSMs addressing this issue (Baroni, Bernardi, & Zamparelli, 2014), including the domains of affixed words (Günther et al., 2018; Marelli & Baroni, 2015) and compounds (Günther & Marelli, 2016, 2018; Marelli, Gagné, & Spalding, 2017). These models

use training procedures to detect the structure guiding the compositional process forming complex word meanings, and can then apply this structure to form new meanings. For example, the affixation model by Marelli and Baroni (2015) uses a regression approach mapping *mummy* to *mummify*, *person* to *personify*, and so on, in order to estimate a function representing the *-ify* affix. Essentially, this function encodes how the *-ify* affix usually affects a word meaning. It can then be applied to new stems, for example to *twitter* in order to create a vector representation for *twitterify*. Similar training approaches have been used to create vector representations for the meaning of novel compounds, such as *honey soup* or *star rock*, from their constituent words (Günther & Marelli, 2016, 2018; Marelli et al., 2017).

To some extent, such approaches are still limited by the words contained in the training set: Clearly identifiable morphological constituents of words have to be repeatedly observed in a corpus so that these models can be trained an applied. However, recent advancements in the domain of natural language processing have addressed this issue by implementing models such as *fastText* (Bojanowski, Grave, Joulin, & Mikolov, 2017) which are trained on sub-lexical units, such as letter n-grams (for example the 3-grams *mou*, *ous*, and *use*), and not only lexical units such as words (*mouse*). In order to represent word meanings, the obtained sub-lexical representations are then used to construct distributional vectors for any combination of these units. In a very recent study (Hendrix, under revision), it has been shown that semantic measures derived from these vectors can indeed be used to predict participants' response times towards nonwords in a lexical decision task.

Thus, even meanings for novel word – new morphologically complex forms as well as completely novel letter combinations – can be represented within a DSM framework, with adequate computational and compositional models at hand. Again, the objection that these are extensions beyond how DSMs are originally conceptualized can be met with the argument that, in order for these systems to work efficiently, all the information must be already available in the statistical regularities in the input captured by DSMs, and can therefore be decoded from the original distributional vectors.

Embodiment and the Role of Non-Linguistic Experience

In the previous section, we have focussed on issues concerning the vector format of meaning representation, and what type of information these vectors can encode. In the following section, we will now turn to the data that is used to create these vectors. This will also lead us to the theoretical debate on embodiment, which has often been framed in opposition to distributional models.

Do DSMs imply that language constitutes the conceptual system?

As described earlier, DSMs are usually trained on large corpora of natural text. Thus, they build their semantic representations solely on the basis of linguistic experience. However, this is surely not representative of human: We have in fact access to substantial amounts of non-linguistic, sensorimotor experience, from perceptual input to our actions in the surrounding world. It is therefore more than plausible that humans heavily rely on this experience as well when building semantic representations. This assumption is at the core of the *embodiment view* of language comprehension (Barsalou, 1999; Glenberg & Kaschak, 2002; Glenberg & Robertson, 2000; Sadoski, 2018; Zwaan & Madden, 2005). Indeed, it has been argued that semantic representations can only be reasonably understood as emerging from and being based on bodily and perceptual experience (Glenberg, 2015).
In this context, DSMs have often been discussed as an opposing theoretical approach, as they rely on purely linguistic input and do not consider sensorimotor experience (Borghesani & Piazza, 2017; Glenberg, Gutierrez, Levin, Japuntich, & Kaschak, 2004; Glenberg & Kaschak, 2002; Glenberg & Robertson, 2000; Munoz-Rubke, Kafadar, & James, 2018; Sadoski, 2018; Simmons, Hamann, Harenski, Hu, & Barsalou, 2008; for an extensive overview on this debate from both theoretical points of view, see De Vega, Glenberg, & Graesser, 2012). Rogers et al. (2004) acknowledge that verbal experience is an important constributor to conceptual knowledge, but since DSMs rely completely on verbal input and are models without any influence of nonverbal experience, they cannot accurately represent our conceptual system (see also Borghesani & Piazza, 2017). Interestingly, it has also been noted that this very property of distributional models renders them ideal candidates to account for meaning representations of abstract concepts that cannot easily be related to sensorimotor experience, such as *equality* or *justice* (Borghi et al., 2017).
We stand on the viewpoint that it is very important in this discussion to distinguish between theoretical assumptions, on the one hand, and practical implementations, on the other hand. While DSMs typically are built entirely from text data, this is in fact only a practical convenience, likely due to the fact that text data is available in huge amounts, and can be very easily and naturally segmented into basic, discrete units (such as words and documents). The theoretical claim of the distributional hypothesis, namely that the contexts in which a word occurs determine its meaning, does not need to be restricted to linguistic contexts alone (although this may have been the case for the original proposal by Harris, 1954). The possibility to include other sources of experience was already discussed in the original article on LSA (Landauer & Dumais, 1997, p. 227): "Indeed, if one judiciously added numerous pictures of scenes with and without rabbits to the

context columns in the encyclopedia corpus matrix, and filled in a handful of appropriate cells in the *rabbit* and *hare* word rows, LSA could easily learn that the words *rabbit* and *hare* go with pictures containing rabbits and not to ones without, and so forth." Similar claims on HAL have been made by Lund and Burgess (1996, p. 29): "We do think a HAL-like model that was sensitive to the same co-occurrences in the natural environment as a human language learner (not just the language stream) would be able to capitalize on this additional information and construct more meaningful representations". Note that such arguments have to assume that the perceptual environment is categorized into discrete units, so that co-occurrences can be counted. Although it is a fundamental psychological insight that humans organize their perceptual input into categories, implementing this into a computational model is certainly not a trivial task. For the sake of this argument, we have to assume that the input to such models is already in a discrete format; see the next section for an overview on how current multimodal DSMs actually integrate perceptual and language information in a single model. We argue that, considering this distinction between theoretical assumptions and practical implementation, it is striking to see that DSMs and (some) models of embodied cognition actually have very much in common. For example, a highly influential and very productive model in the embodied tradition is the *experiential trace model* by Zwaan and Madden (2005). This model assumes that all mental representations are formed from experiential traces, which can be divided into sensorimotor and linguistic traces. Critically, these traces are interconnected, and these interconnections are assumed to arise from *co-occurrences*: Co-occurrences between sensorimotor traces (for example, seeing a dog together with a leash), co-occurrences between sensorimotor and linguistic traces (seeing a dog while hearing the word *dog*), and co-occurrences between linguistic traces (hearing the words *dog* and *leash* in the same linguistic context). Zwaan and Madden (2005) explicitly mention that DSMs such as LSA can be used to model the latter kind, i.e. linguistic co-occurrences. However, when dropping the implicit assumption that DSMs are inherently language-centric – which is not an actual theoretical assumption of these models – and accepting that they can also take sensorimotor experience as *context* (see also Hasson, Egidi, Marelli, & Willems, 2018 for a similar argument from a neurobiological point of view), then the experiential trace model can in fact be itself identified as a DSM.

Are DSMs based on non-linguistic experience actually implemented?

The view that the exclusion of sensorimotor experience might be a practical rather than a theoretical issue has been acknowledged by some critics of DSMs (see Glenberg et al., 2004; Glenberg & Robertson, 2000). However, one of their points of criticism against solutions such as those proposed by Landauer and Dumais (1997) or Lund and Burgess (1996) is that these are not actually implemented in the models. In fact, similar argumentations can be found within the DSM literature itself, as stated by Sahlgren (2006, p. 135): "In order for the word-space model to qualify as a model of human semantic processing, it needs to reach beyond the linguistic frontier into the realms of the extralinguistic world, and to include extralinguistic context in its representation. I do not believe that this is impossible in principle, although I do believe that it would require a radical innovation in how we define and use context for accumulating context vectors. Until that breakthrough, we should be wary about claims for cognitive plausibility." Therefore, the claim that "the central deficiency of AI-based theories is the lack of a convincing account of comprehension and meaning" (Sadoski, 2018, p. 337) is perpetuated in the literature: In a very recent review article, Sadoski (2018) concludes that "disembodied computer vectors of word relationships could not comprehend situations that would be quite simple even for young children" (p. 337), explicitly referring to the arguments and study by Glenberg and Robertson (2000).

Yet, following up on the initial objections, there has been a surge in recent years in research on *multimodal DSMs* (e.g. Andrews, Vigliocco, & Vinson, 2009; Bruni et al., 2014; Kiela & Clark, 2015; Kiela, Bulat, & Clark, 2015; Lazaridou, Bruni, & Baroni, 2014; Lazaridou, Pham, & Baroni, 2015; Lopopolo & Miltenburg, 2015; Sadeghi, McClelland, & Hoffman, 2015; see also Lee & Seung, 1999). However, also in very recent presentations of DSMs as purely language-centric models, these newer developments are often not considered (Sadoski, 2018; Munoz-Rubke et al., 2018).

Multimodal DSMs are explicitly designed to incorporate non-linguistic (experiential and perceptual) and linguistic data in order to create distributional vectors. From a theoretical point of view, they are therefore in line with models of embodied cognition such as the experiential trace model cited above (Zwaan & Madden, 2005). For instance, Howell, Jankowicz, and Becker (2005) showed, using a neural network training technique, that vectors trained on text corpora and ratings on sensorimotor features collected from participants performed better than vectors trained on text corpora alone. Similarly, Andrews et al. (2009) use a Bayesian approach to integrate traditional distributional vectors with speaker-generated feature norms into combined semantic representations (see Steyvers, 2010, for a similar approach). Using machine learning algorithms, Bruni et al. (2014) extract vectors representing visual features directly from images, and then merge these with traditional distributional vectors for the associated image labels in order to construct more comprehensive semantic representations. The model by Lazaridou et al. (2014) learns

a mapping function between visual and textual vectors from training examples, which can then be applied to instances outside the training set. Lazaridou et al. (2015) use an algorithm similar to the *word2vec* model (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013) to predict linguistic contexts simultaneously with visual features from words in order to construct distributional vectors (the words were used as labels for the images, and visual vectors representing the visual features were extracted automatically from these images). In a similar vein, Sadeghi et al. (2015) applied an LSA-like algorithm on a corpus of labelled objects (corresponding to the words in LSA) in visual scenes (corresponding to the documents in LSA) to derive distributional vectors for the object labels.

Perhaps more surprisingly, in other perceptual domains models have been designed that incorporate sound data (Kiela & Clark, 2015; Lopopolo & Miltenburg, 2015) or even olfactory data (Kiela et al., 2015) into distributional vectors, following the principle methods employed in the visual domain as outlined above. Interestingly, these multimodal models have been shown to outperform traditional DSMs based only on text data in a variety of tasks, for example in predicting human similarity ratings (Bruni et al., 2014). They can also overcome problems associated to text-based DSMs, which for example tend to undervalue basic perceptual information.

That said, it is apparent that these models rely largely on visual data extracted from images, mainly because of the availability of large annotated image data bases and methods to extract vectors from image data (see however Kiela & Clark, 2015; Kiela et al., 2015; Lopopolo & Miltenburg, 2015). The scope of the embodiment view obviously goes far beyond this. However, it is important to highlight again that the main issue is collecting adequate training data, and it is thus not directly related to the model architectures or learning algorithms incorporated in DSMs. In other words, the issue is practical rather than theoretical, as already discussed in the previous section. In this context, the already existing and fully implemented models described here show that non-linguistic data can be incorporated to build distributional vectors, improving their quality. This is an important first step in the direction of more realistic DSMs using all kinds of sensorimotor and linguistic experience available to humans.

Do language-based DSMs have access to sensorimotor information?

Although multimodal DSMs have become available recently, the vast majority of research employing DSMs has used traditional models trained on language corpora. In many cases, these have been used to model conceptual knowledge. But how can this be plausible, given that these DSMs only have access to language data, while our conceptual knowledge heavily relies on sensorimotor information?

Due to this supposed limitation, it has been argued that text-based DSMs can only serve as models of lexical semantics (and are therefore restricted to the linguistic or engineering domain), but not as models of the human conceptual system (Barsalou, 1999; French & Labiouse, 2002; Glenberg & Robertson, 2000; Glenberg et al., 2004; Sadoski, 2018). This argumentation relies on the assumption that the language we use is independent from the world we live in, and from our sensorimotor experience. In reality, however, one of the primary functions of language is to confer to others information about the world we perceive and act in. Therefore, a substantial amount of sensorimotor information is actually incorporated in language - for example, the fact that some food is tasty, while other things that can be eaten are far from being tasty.

The idea that sensorimotor information and language are intertwined has been put forward by Louwerse in the *symbol interdependency hypothesis* (e.g. Louwerse, 2011; Louwerse & Zwaan, 2009). This hypothesis receives support from the observation that distributional vectors actually encode surprising amounts of information about the surrounding world: For example, Louwerse and Zwaan (2009) applied a multidimensional scaling technique to the similarities between distributional vectors for city names, and projected them onto a two-dimensional space. They found that the coordinates of the cities within this space were correlated with the actual geographical positions of these cities, in the real world (Louwerse & Zwaan, 2009; Recchia & Louwerse, 2016) as well as in fictional worlds such as Lord of the Ring's Middle Earth (Louwerse & Benesh, 2012). Other studies have shown that the similarity structure between distributional vectors also reflects other real-world similarity structures, such as the relative position of the days of the week or months of the year (Louwerse, Raisig, Tillman, & Hutchinson, 2015). Importantly for the argument presented here, studies testing the symbol interdependency hypothesis have also found that distributional vectors encode the vertical location of objects in the world (Hutchinson & Louwerse, 2013), perceptual modalities (*visual*, *auditory*, *olfactory*, *gustatory* and *haptic*; Louwerse & Connell, 2011), affective dimensions (*dominance*, *valence* and *arousal*; Hollis & Westbury, 2016), the typical spatial organization of objects, or the sensibility of performing certain actions with specific objects (Louwerse, 2011). Thus, distributional vectors do not only encode what would generally be considered world knowledge, but also certain sensorimotor aspects of our conceptual systems that are usually explained by referring to embodied or grounded approaches of meaning.

In summary, because language is not independent from the world it is used in, but used to communicate about said world, the structure of semantic relations within language tends to

reflect the structure of the outside world (see also Andrews et al., 2009; Johns & Jones, 2012; Hoffman et al., 2018; Riordan & Jones, 2011). Since distributional vectors are heavily influenced by statistical regularities of language, they also incorporate substantial amounts of sensorimotor experience. Note however that, although there is a correspondence between information expressed in language and the physical world, we (and the literature cited above) do not claim this correspondence to be perfect. Thus, while some aspects of meaning can be obtained from both sources, others are only available from either of them (Bruni et al., 2014).

Open Issues

In the previous sections, we have discussed a wide range of arguments that are frequently directed at DSMs. We have outlined how these arguments have already been or can successfully be addressed within the DSM framework, and thus are not in-principle arguments against the validity of DSMs as models of human semantic representations. However, there are still open issues that need to be addressed. In the following sections, we will outline some of these issues as potentially paramount topics for future research on the cognitive validity of DSMs.

Models and Training Data

In a number of arguments throughout the article, we have pointed out that a series of initial criticisms against DSMs may be related to the data on which the models are trained, rather than to the models themselves. It is important to point out again that the distinction between the model architecture on the one hand and the training data on the other hand is a critical one, and that highly valid arguments against one aspect cannot readily be taken as valid arguments against the other.

In most instances, we have argued that objections raised against DSMs are in fact arguments against the training data: This especially concerns the issue of multimodality, but also points such as the cognitive plausibility of the models and the access to perceptual information through language. While research on DSMs has been constantly striving to improve the quality and adequacy of the training data, this issue is far from solved. Obtaining optimal training data therefore is an open issue to be addressed in future research.

In this context, we argue that several points should be considered. A common approach in machine learning and natural language processing to obtain higher-quality training data is increasing the amount of data itself, which usually improves the quality of the empirical results (Recchia & Jones, 2009). Having more training data makes the model algorithms more robust against statistical noise, and also increases the representativity of the training data for a population of speakers, as usually investigated in empirical studies. The improvement of model performance with corpus size has also been suggested to determine better inferences about perceptual representations: These are based on redundancies between the perceptual world and language (see Louwerse, 2011), which can be more easily detected with more language experience available to the model (Johns & Jones, 2012).

However, increasing the amount of training data cannot be the whole answer to the question. For once, not only the size, but also the structure of the training data (i.e. the type of data included) is a critical factor for its adequacy. For example, Herdağdelen and Marelli (2017) have demonstrated that word frequencies derived from social media corpora – which can be assumed to be close to actual human language experience – give better results for psycholinguistic studies than other general-purpose corpora. Thus, optimal training data for DSMs should be as representative of human experience as possible, which would also be desirable in order to raise the cognitive plausibility of these models.

Another issue to be considered in this context is the fact that DSMs are usually set up to model semantic representations on the population level rather than the individual level. However, this approach is inherently inadequate to account for individual differences between speakers in a DSM framework (see Schmidtke, Van Dyke, & Kuperman, 2018, for results indicating that typical DSMs are only adequate for highly literate persons). In an "ideal" world (from a research perspective), it would be highly valuable to build semantic spaces for individual speakers based on their idiosyncratic experience (for an impressive study collecting three years of video material of a researcher's child to study its language acquisition, see Roy et al., 2006). In this context, Johns, Jones, and Mewhort (2016) proposed a method to sample documents from a large multi-source corpus in order to estimate the sub-corpus best representing the language experience of a speaker group or individual speakers, in that it best captures standard behavioural effects such as the word-frequency effect (Brysbaert, Mandera, & Keuleers, 2018) in these participants' performance. While relying on such behavioural "training data" might not always be practicable or feasible, their results suggest some rule-of-thumb approaches to the selection of adequate sub-corpora: For example, the estimated sub-corpus representing the experience of younger participants contained far more young-adult books than that of older participants. In principle, the technical instruments for actual large-scale individual data collection are readily available in the age of the internet and social media, for data experienced as well as produced by speakers; the question is more if and how such a line of research should be pursued, given the severe ethical issues it raises.

Finally, an issue where the technical solution is less advanced is the collection of data that include actual multimodal experience. At the moment, multimodal DSMs largely rely on annotated data sets of sensorimotor information (such as im-

ages or sound files), which are then integrated with information collected from linguistic corpora. However, either are usually obtained from independent sources of data: The training corpora typically don't consist of text used in the context of the images (or other sensorimotor information). This is a crucial shortcoming for some embodied theories of meaning, where the notion of co-occurrence between linguistic and sensorimotor traces of experience is central (Zwaan & Madden, 2005). Thus, an ideal data base for truly multimodal DSMs would consist of a large collection of actual experience, where co-occurring entities from different modalities and sources (sensorimotor and linguistic) are simultaneously encoded. We acknowledge that this is not a trivial task. A possible starting point that seems to be technologically feasible in the near future might be the decoding of video material, employing techniques developed in artificial intelligence research, such as visual object recognition and speech recognition.

A Unified Distributional Model?

In this article, we have addressed a range of issues that are at times associated with DSMs. In many cases, we have argued that most of these issues are not inherent to DSMs per se, and can be solved with appropriate model architectures: Incremental models such as BEAGLE or *word2vec* don't rely on implausible learning assumptions, multimodal DSMs are not dependent on language experience only, and Topic Models don't produce arbitrary, non-interpretable semantic dimensions. All these approaches constitute valuable proofs of concept for the possibilities within the general DSM approach. However, it is also obvious that, at this point, we have different models designed to address different issues. From a theoretical point of view, this is unsatisfying (Rogers & Wolmetz, 2016): If DSMs are to serve as a plausible model of semantic representation, there should be a unified DSM architecture which can address all issues simultaneously. A similar argument holds for the data on which the models are trained (as discussed in the previous section): If different training data is needed for different models or to address different phenomena, this clearly comes at the cost of generalizability.

A necessary condition for the implementation of a unified DSM is that the different approaches are not mutually exclusive. In the case of the three examples described above, the model architectures should be compatible with one another: Incremental models rely on specific learning algorithms that processes chunks of data successively; multimodal DSMs rely on different input channels for different types of data, and need a mechanism to merge these channels; and Topic Models represent words as distributions over topics (and vice versa). For all pairwise combinations, these approaches have been shown to be compatible: The multimodal skip-gram model by Lazaridou et al. (2017) incorporates an incremental learning algorithm as implemented in the *word2vec* model (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013). Steyvers (2010) has demonstrated that perceptual data – here in the form of feature norms – can be integrated in Topic Models alongside textual data (see also Andrews et al., 2009). Finally, Topic Models can also be built incrementally (AlSumait, Barbará, & Domeniconi, 2008). Thus, it should in principle be possible to design a mechanism that incrementally builds a Topic Model from different kinds of input data.

We therefore believe that a unified DSM can be designed, and advocate this to be an objective of future research. As in any scientific field, theory and model development within the DSM framework would benefit from a cumulative and nested approach (compare Jacobs & Grainger, 1994). Under this approach, an existing model should be adjusted and extended in order to incorporate new mechanisms and to account for new results, rather than crafting novel architectures that don't consider previous insights.

A Variety of Tasks for Semantic Memory

Any model claiming to be a general-level model of semantic memory has to satisfy at least two criteria: (1) It has to successfully predict human behaviour across the wide range of tasks in which semantic memory plays a role, including behavioural data it was not designed for and not trained on; (2) its performance should be as close to actual human performance in these tasks as possible. Thus, a model should not fail in tasks that humans are able to perform, nor should it vastly outperforms humans (which computational models at times tend to do).

In most respects, DSMs serve these criteria quite well. They were conceived as models of semantic memory in the late 90s, without being tailored for any specific task. Follow-up research has then subsequently shown that they account quite well for human behaviour across a variety of tasks without being explicitly designed for them (e.g. Bullinaria & Levy, 2012; Jones et al., 2006; Louwerse, 2011; Mandera et al., 2017). Thus, the necessary criteria for serving as general-level models of semantic memory are satisfied. Yet, there are still open issues. First, DSMs don't perform equally well in all tasks. For example, they tend to have very high performances in relatively explicit meaning judgement tasks, such as synonym tests (Bullinaria & Levy, 2012), similarity judgements (Bruni et al., 2014), or categorization tasks (Baroni & Lenci, 2010). On the other hand, their performance is somewhat lower in other tasks, such as semantic priming (Hutchison, Balota, Cortese, & Watson, 2008; Mandera et al., 2017) or free associations (Kennett, Levi, Anaki, & Faust, 2017; Nematzadeh et al., 2017).[6]

[6]Note that these studies did not consider asymmetrical similarity measures as proposed by Kintsch (2014), which might be more adequate for free associations.

Additionally, not all DSMs perform equally well across all tasks. For example, while the recent *word2vec* model tend to outperform more traditional models across almost all tasks (such as similarity judgements, categorizations, synonym tests, semantic priming; Baroni, Dinu, & Kruszewski, 2014; Mandera et al., 2017; Pereira et al., 2016), it is in turn outperformed by Topic Models in predicting free associations (Nematzadeh et al., 2017). Additionally, DSMs usually have a number of free parameters (for example number of dimensions, or context size), which also have differential effects on their performance for different tasks (Bullinaria & Levy, 2007, 2012).

Ideally, it would be highly desirable to have one single DSM with a fixed set of parameters that is able to predict human behaviour across the whole range of semantic memory-related tasks with a high accuracy. Obviously, this is a very ambitious objective – at this point, we cannot predict whether it is bound to succeed, which ultimately remains an empirical question. However, we again want to advocate to converge on a unified DSM instead of crafting many different models specifically designed to shine in exactly one task. While this is a valid strategy for language engineering, it is only of limited helpfulness in building a unified model and theory of the human semantic memory.

Learning Beyond Co-occurrence Patterns

By definition, DSMs define learning as observing and abstracting the distributional patterns of entities over contexts (or co-occurrence patterns) in the vast amount of experience available to us. In the earlier sections, we discussed that this is not limited to linguistic entities, but can instead be applied to all different kinds of entities, including the general domains of perception and action. By applying this broader definition, DSMs can seamlessly be integrated with embodied and grounded models of cognition, which also rely heavily on the notion of co-occurrence (Zwaan & Madden, 2005).

One main question remains open: Can all kinds of human learning be broken down in terms of distributional learning? For learning mechanisms that are not necessarily defined in terms of distributional patterns, possible solutions can be imagined: For example, evaluative conditioning (learning to like or dislike stimuli De Houwer, Thomas, & Baeyens, 2001; Levey & Martin, 1975) can be conceptualized as distributional patterns of stimuli over evaluative psychological states. Admittedly, this operationalization might be difficult to actually implement in a real-world scenario. However, on a conceptual level, we argue that all types of learning that can be traced back to a process of classical conditioning or association learning (as is the case for, for example, evaluative conditioning De Houwer et al., 2001) can be translated into distributional patterns.

At the same time, there are certainly cases where conceptualizations of learning scenarios as co-occurrence patterns reach their limit. A possible example is one-shot learning, which describes how humans, and especially children, are sometimes able to learn new concepts from just a single exposition (Landau, Smith, & Jones, 1988). This seems out of the scope of distributional models, which usually learn their representations from vast amounts of data. This is rooted in the assumption that DSMs learn semantic representations from the repetition and abstraction of co-occurrence patterns. Thus, because they are explicitly modelled to transcend episodic memory, they cannot easily model phenomena associated to it. Despite there being computational models that successfully model one-shot learning (Lake, Salakhutdinov, & Tenenbaum, 2015), these rely on a different model architecture than current DSMs, and it is at the current point unclear whether the two approaches can be combined into a general-purpose model architecture (although models such as *word2vec* seem to be potential candidates for this task, and studies relying on such architectures have shown that they can learn new concepts from minimal exposure; Lazaridou et al., 2017).

Furthermore, current DSMs are heavily focussed towards the stimulus level of learning. In most cases, the entities as well as the contexts in which they occur are defined in terms of stimuli, such as words, documents, or visual features in multi-modal models. There is, however, a large body of psychological literature demonstrating that learning is more than just passively encoding and observing a stream of stimuli: Factors such as attention (Gottlieb, 2012; Grossberg, 1999), motivation (Dweck, 1986), emotion (Hascher, 2010), or personality traits (Corr, Pickering, & Gray, 1995), amongst many others, all can influence learning and the way that we organize the experience we make. Although research on DSMs has undertaken first steps in the direction of integrating these factors into the models (see Ling et al., 2015, for a *word2vec*-based model implementing some basic – not necessarily psychologically motivated – attentional mechanisms), this line of work is still in its infancy, and requires considerable amounts of additional research.

Another critical issue in the DSM approach to learning is that the learner is essentially conceptualized as a passive recipient of experience, an observer. While this is certainly an important aspect of language acquisition, it does not consider that humans are also active learners that interact with the world and other humans. It has been argued that such interactions with the world and with others play a central role in constituting our semantic representations (e.g. Glenberg & Kaschak, 2002; Glenberg, 2015). This view receives strong empirical support from recent studies in artificial intelligence, where artificial agents learn semantic representations for initially meaningless symbols through communication with each other, updating them based on whether a communicative act was successful or not (e.g. Spranger, 2012; Spranger, Pauw, Loetzsch, & Steels, 2012). From

such a perspective, the co-occurrence of words follows from the agents' communication and the structure of their environment, rather than constituting the basis of their semantic representations (cf. Louwerse, 2011). This corresponds to a weak version of the distributional hypothesis, assuming that a word's distribution follows from its meaning (Lenci, 2008). Surely, the human acquisition of semantic representation entails both processes: The construction of semantic representations based on the distributional structure of entities in their environment (passive observation), as well as the interaction with the environment and communication about it (active engagement). Investigating the relative role of these processes, also with respect to the time course of language acquisition, remains a central, open issue for future research.

Conclusion

Mental representations of the world are a highly powerful cognitive tool that enable us to successfully navigate in and properly interact with our environment, and help us to structure the endless and overwhelming stream of information to which we are continuously exposed. However, every single one of us experiences a highly idiosyncratic environment, specific to time, location, culture, and the people surrounding us, and also subject to constant change. Mental representations of the world are hence most useful if they are shaped through these specific experiences we make. In this sense, learning consists in perpetually tuning our cognitive system to the environment in which we live, and the experience we make (Ramscar, Hendrix, Love, & Baayen, 2013). And, indeed, humans excel in capturing and learning statistical regularities and patterns in their environment across many different domains (Perruchet & Pacton, 2006; Saffran, 2003). As stated by Anderson and Schooler (1991, p. 404): "human memory mirrors, with a remarkable degree of fidelity, the structure that exists in the environment".

DSMs seem to take on this function. Taking together the literature reviewed in the present article, a common overarching theme is that statistical regularities and redundancies – here, in the form of co-occurrences and distributional patterns – entail information that can give rise to rich mental representations, and to cognitive phenomena usually described through high-level concepts (see Westbury, 2016). In order to build these representations, we "just" need to extract this statistical information from our experience through some powerful learning mechanisms. Somewhat rather surprisingly given the complexity of the stream of information we are exposed to, it seems that such powerful mechanisms can actually be rather simple: A common element between all the approaches reviewed here is that they focus on the informative relation between two units – if the presence/absence of one stimulus is a valid cue for the presence/absence of another stimulus – rather than their simple co-presence (Rescorla & Wagner, 1972). Further, these models don't only rely on local co-occurrences, but also consider global co-occurrence patterns in order to structure the incoming information (Landauer & Dumais, 1997). Just think of two bus drivers on your line which you probably never see together, but always in the same context: the same bus, the same time, the same route. Even though you only see one of them at any given time, your representations of them will nevertheless be highly similar.

How does language come into this picture? On the one hand, language itself provides stimuli we experience on a very regular basis, and which occur in the context of other linguistic and non-linguistic stimuli (as extensively discussed in the last section of the article; see also Zannino, Caltagirone, & Carlesimo, 2015). As a result, the linguistic system is ideally suited as a learning environment. On the other hand, language serves a crucial function as "experience by proxy" (Johnson-Laird, 1983), allowing us to transfer experience across persons, across space, and even across time. We can therefore build mental representations of the world that are potentially informed by the experience of an entire species, and truly stand on the shoulders of giants who help us navigate through this world (Harari, 2014).

References

AlSumait, L., Barbará, D., & Domeniconi, C. (2008). On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. In *Eighth IEEE International Conference on Data Mining, 2008 (ICDM'08).* (pp. 3–12).

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, *2*, 396–408.

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, *116*, 463–498.

Annesi, P., Croce, D., & Basili, R. (2013). Towards compositional tree kernels. In *Proceedings of the Joint Symposium on Semantic Processing. Textual Inference and Structures in Corpora* (pp. 15–23). Trento, Italy.

Arora, S., Ge, R., & Moitra, A. (2012). Learning topic models – Going beyond SVD. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium* (pp. 1–10).

Baroni, M., Bernardi, R., & Zamparelli, R. (2014). Frege in space: A program for compositional distributional semantics. *Linguistic Issues in Language Technologies*, *9*(6), 5–110.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of ACL 2014* (pp. 238–247). East Stroudsburg, PA: ACL.

Baroni, M., & Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, *36*, 673–721.

Baroni, M., Murphy, B., Barbu, E., & Poesio, M. (2010). Strudel: A Corpus-Based Semantic Model Based on Properties and Types. *Cognitive Science*, *34*, 222–254.

Baroni, M., & Zamparelli, R. (2010). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the Conference on Empirical*

Methods in Natural Language Processing (pp. 1183–1193). East Stroudsberg, PA: ACL.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*, 637–660.

Bhatia, S. (2017a). Associative judgment and vector space semantics. *Psychological Review*, *124*, 1–20.

Bhatia, S. (2017b). The semantic representation of prejudice and stereotypes. *Cognition*, *164*, 46–60.

Biemann, C., & Riedl, M. (2013). Text: Now in 2d! a framework for lexical expansion with contextual similarity. *Journal of Language Modelling*, *1*, 55–95.

Bierwisch, M. (2011). Semantic features and primes. In C. Maienborn, K. von Heusinger, & P. Portner (Eds.), *Semantics: An International Handbook of Natural Language Meaning* (Vol. 1, p. 322-357). Berlin, Germany: de Gruyter.

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

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Boleda, G., Padó, S., & Utt, J. (2012). Regular polysemy: A distributional model. In *Proceedings of *SEM* (pp. 151–160). Montreal, Canada: ACL.

Borghesani, V., & Piazza, M. (2017). The neuro-cognitive representations of symbols: the case of concrete words. *Neuropsychologia*, *105*, 4–17.

Borghi, A. M., Binkofski, F., Castelfranchi, C., Cimatti, F., Scorolli, C., & Tummolini, L. (2017). The challenge of abstract concepts. *Psychological Bulletin*, *143*, 263–292.

Borghi, A. M., Glenberg, A. M., & Kaschak, M. P. (2004). Putting words in perspective. *Memory & Cognition*, *32*, 863–873.

Bruni, E., Tran, N.-K., & Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, *49*, 1-47.

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science*, *27*, 45–50.

Buitelaar, P., Cimiano, P., & Magnini, B. (2005). Ontology learning from text: An overview. In *Ontology learning from text: Methods, evaluation and applications* (pp. 3–12). Amsterdam, The Netherlands: IOS Press.

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

Bullinaria, J. A., & Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD. *Behavior Research Methods*, *44*, 890–907.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Camacho-Collados, J., & Pilehvar, T. (2018). From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, *63*, 743–788.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems (NIPS) 2009* (Vol. 22, pp. 288–296).

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*, 22–29.

Cohen, T., & Widdows, D. (2009). Empirical distributional semantics: Methods and biomedical applications. *Journal of Biomedical Informatics*, *42*, 390–405.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*, 207–428.

Corr, P. J., Pickering, A. D., & Gray, J. A. (1995). Personality and reinforcement in associative and instrumental learning. *Personality and Individual Differences*, *19*, 47–71.

De Houwer, J., Thomas, S., & Baeyens, F. (2001). Association learning of likes and dislikes: A review of 25 years of research on human evaluative conditioning. *Psychological Bulletin*, *127*, 853–869.

Dennis, S. (2007). How to Use the LSA Website. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 57–70). Mahwah, NJ: Erlbaum.

De Vega, M., Glenberg, A., & Graesser, A. (2012). *Symbols and embodiment: Debates on meaning and cognition*. Oxford University Press.

Dinu, G., & Lapata, M. (2010). Measuring distributional similarity in context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1162–1172). Cambridge, MA.

Dinu, G., Pham, N., & Baroni, M. (2013). DISSECT: DIStributional SEmantics Composition Toolkit. In *Proceedings of the System Demonstrations of ACL 2013 (51st Annual Meeting of the Association for Computational Linguistics)* (pp. 31–36). East Stroudsburg, PA: ACL.

Dreyer, F. R., & Pulvermüller, F. (2018). Abstract semantics in the motor system? An event-related fMRI study on passive reading of semantic word categories carrying abstract emotional and mental meaning. *Cortex*, *100*, 52–70.

Dudschig, C., Maienborn, C., & Kaup, B. (2016). Is there a difference between stripy journeys and stripy ladybirds? The N400 response to semantic and world-knowledge violations during sentence processing. *Brain and Cognition*, *103*, 38–49.

Durda, K., Buchanan, L., & Caron, R. (2009). Grounding co-occurrence: Identifying features in a lexical co-occurrence model of semantic memory. *Behavior Research Methods*, *41*, 1210–1223.

Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist*, *41*, 1040–1048.

Elvevåg, B., Foltz, P. W., Weinberger, D. R., & Goldberg, T. E. (2007). Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. *Schizophrenia Research*, *93*, 304–316.

Evangelopoulos, N., Zhang, X., & Prybutok, V. R. (2012). Latent semantic analysis: five methodological recommendations. *European Journal of Information Systems*, *21*, 70–86.

Firth, J. R. (1957). *Papers in linguistics, 1934–1951*. Oxford, UK: Oxford University Press.

French, R. M., & Labiouse, C. (2002). Four problems with extracting human semantics from large text corpora. In *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society* (pp. 316–322). Mahwah, NJ: Erlbaum.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological Review*, *103*, 650–669.

Glenberg, A. M. (2015). Few believe the world is flat: How embodiment is changing the scientific understanding of cognition. *Canadian Journal of Experimental Psychology*, *69*, 165–171.

Glenberg, A. M., Gutierrez, T., Levin, J. R., Japuntich, S., & Kaschak, M. P. (2004). Activity and imagined activity can enhance young children's reading comprehension. *Journal of Educational Psychology*, *96*, 424–436.

Glenberg, A. M., & Kaschak, M. P. (2002). Grounding language in action. *Psychonomic Bulletin & Review*, *9*, 558–565.

Glenberg, A. M., & Mehta, S. (2008). Constraints on covariation: It's not meaning. *Italian Journal of Linguistics*, *20*, 241–264.

Glenberg, A. M., & Robertson, D. A. (2000). Symbol Grounding and Meaning: A Comparison of High-Dimensional and Embodies Theories of Meaning. *Journal of Memory and Language*, *43*, 379–401.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, *117*(3), 227.

Goldberg, R. F., Perfetti, C. A., & Schneider, W. (2006). Perceptual knowledge retrieval activates sensory brain regions. *Journal of Neuroscience*, *26*, 4917–4921.

Gottlieb, J. (2012). Attention, learning, and the value of information. *Neuron*, *76*, 281–295.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology*, *74*, 1464–1480.

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

Grossberg, S. (1999). The link between brain learning, attention, and consciousness. *Consciousness and Cognition*, *8*, 1–44.

Günther, F., Dudschig, C., & Kaup, B. (2015). LSAfun–An R package for computations based on Latent Semantic Analysis. *Behavior Research Methods*, *47*, 930–944.

Günther, F., Dudschig, C., & Kaup, B. (2016). Latent semantic analysis cosines as a cognitive similarity measure: Evidence from priming studies. *The Quarterly Journal of Experimental Psychology*, *69*, 626–653.

Günther, F., & Marelli, M. (2016). Understanding Karma Police: The Perceived Plausibility of Noun Compounds as Predicted by Distributional Models of Semantic Representation. *PLOS ONE*, *11*(10). doi: 10.1371/journal.pone.0163200

Günther, F., & Marelli, M. (2018). Enter Sandman: Compound Processing and Semantic Transparency in a Compositional Perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Advance online publication. doi: 10.1037/xlm0000677

Günther, F., Smolka, E., & Marelli, M. (2018). 'Understanding' differs between English and German: Capturing Systematic Language Differences of Complex Words. *Cortex*, Advance online publication. doi: 10.1016/j.cortex.2018.09.007

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science*, *304*, 438–441.

Hansson, K., Bååth, R., Löhndorf, S., Sahlén, B., & Sikström, S. (2016). Quantifying semantic linguistic maturity in children. *Journal of Psycholinguistic Research*, *45*, 1183–1199.

Harari, Y. N. (2014). The Tree of Knowledge. In *Sapiens: A brief history of humankind* (pp. 20–40). London, UK: Random House.

Harris, Z. (1954). Distributional Structure. *Word*, *10*, 146–162.

Hascher, T. (2010). Learning and emotion: perspectives for theory and research. *European Educational Research Journal*, *9*, 13–28.

Hasson, U., Egidi, G., Marelli, M., & Willems, R. M. (2018). Grounding the neurobiology of language in first principles: The necessity of non-language-centric explanations for language comprehension. *Cognition*, *180*, 135–157.

Heider, E. R., & Olivier, D. C. (1972). The structure of the color space in naming and memory for two languages. *Cognitive Psychology*, *3*, 337–354.

Hendrix, P. (under revision). A word or two about nonwords: nonword frequency and semantic neighborhood density effects in the lexical decision task.

Herdağdelen, A., & Marelli, M. (2017). Social media and language processing: How facebook and twitter provide the best frequency estimates for studying word recognition. *Cognitive Science*, *41*, 976–995.

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.

Hoffman, P., McClelland, J. L., & Lambon Ralph, M. A. (2018). Concepts, control, and context: A connectionist account of normal and disordered semantic cognition. *Psychological Review*, *125*, 293–328.

Hollis, G. (2017). Estimating the average need of semantic knowledge from distributional semantic models. *Memory & Cognition*, *45*, 1350–1370.

Hollis, G., & Westbury, C. (2016). The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic Bulletin & Review*, *23*, 1744–1756.

Howell, S. R., Jankowicz, D., & Becker, S. (2005). A model of grounded language acquisition: Sensorimotor features improve lexical and grammatical learning. *Journal of Memory and Language*, *53*, 258–276.

Hutchinson, S., & Louwerse, M. M. (2013). What's up can be explained by language statistics. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society* (pp. 2596–2601). Berlin, Germany.

Hutchison, K. A., Balota, D. A., Cortese, M., & Watson, J. M. (2008). Predicting semantic priming at the item level. *The Quarterly Journal of Experimental Psychology*, *61*, 1036–1066.

Jackendoff, R. (2003). Précis of foundations of language: brain, meaning, grammar, evolution. *Behavioral and Brain Sciences*, *26*, 651–665.

Jacobs, A. M., & Grainger, J. (1994). Models of visual word recognition: Sampling the state of the art. *Journal of Experimental Psychology: Human Perception and Performance*, *20*, 1311–1334.

Jenkins, J. J. (1954). Transitional organization: Association techniques. In C. E. Osgood & T. A. Sebeok (Eds.), *Psycholinguis-*

tics. A Survey of Theory and Research Problems (p. 112-118). Bloomington, IN: Indiana University Press.

Johns, B. T., Jones, M., & Mewhort, D. J. (2016). Experience as a free parameter in the cognitive modeling of language. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 2291–2296). Austin, TX: Cognitive Science Society.

Johns, B. T., & Jones, M. N. (2012). Perceptual Inference Through Global Lexical Similarity. *Topics in Cognitive Science*, *4*, 103–120.

Johnson, K. (2008). An overview of lexical semantics. *Philosophy Compass*, *3*, 119–134.

Johnson-Laird, P. N. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge, MA: Harvard University Press.

Jones, M. N., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, *55*, 534–552.

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

Jones, M. N., Willits, J., & Dennis, S. (2015). Models of semantic memory. In J. Busemeyer, Z. Wang, J. Townsend, & A. Eidels (Eds.), *Oxford Handbook of Mathematical and Computational Psychology* (pp. 232–254). New York, NY: Oxford University Press.

Jorge-Botana, G., León, J. A., Olmos, R., & Hassan-Montero, Y. (2010). Visualizing polysemy using LSA and the predication algorithm. *Journal of the Association for Information Science and Technology*, *61*, 1706–1724.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*, 237–251.

Kaschak, M. P., & Glenberg, A. M. (2000). Constructing meaning: The role of affordances and grammatical constructions in sentence comprehension. *Journal of Memory and Language*, *43*, 508–529.

Kennett, Y. N., Levi, E., Anaki, D., & Faust, M. (2017). The semantic distance task: Quantifying semantic distance with semantic network path length. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*, 1470–1489.

Kiela, D., Bulat, L., & Clark, S. (2015). Grounding semantics in olfactory perception. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (pp. 231–236). Beijing, China: ACL.

Kiela, D., & Clark, S. (2015). Multi-and cross-modal semantics beyond vision: Grounding in auditory perception. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)* (pp. 2461–2470). Lisbon, Portugal: ACL.

Kintsch, W. (1988). The use of knowledge in discourse processing: A construction-integration model. *Psychological Review*, *95*, 163–182.

Kintsch, W. (2000). Metaphor comprehension: A computational theory. *Psychonomic Bulletin & Review*, *7*, 257–266.

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

Kintsch, W. (2014). Similarity as a Function of Semantic Distance and Amount of Knowledge. *Psychological Review*, *121*, 559–561.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition*, *83*, B35–B42.

Kowialiewski, B., & Majerus, S. (2018). The non-strategic nature of linguistic long-term memory effects in verbal short-term memory. *Journal of Memory and Language*, *101*, 64–83.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*, 1332–1338.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, *3*, 299–321.

Landauer, T. K. (2007). LSA as a Theory of Meaning. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 3–34). Mahwah, NJ: Erlbaum.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes*, *25*, 259–284.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (2007). Handbook of Latent Semantic Analysis. Mahwah, NJ: Erlbaum.

Lang, E., & Maienborn, C. (2011). Two-level semantics: Semantic form and conceptual structure. In C. Maienborn, K. von Heusinger, & P. Portner (Eds.), *Semantics: An International Handbook of Natural Language Meaning* (Vol. 1, pp. 709–740). Berlin, Germany: de Gruyter.

Lau, M. C., Goh, W. D., & Yap, M. J. (2018). An item-level analysis of lexical-semantic effects in free recall and recognition memory using the megastudy approach. *The Quarterly Journal of Experimental Psychology*, Advance online publication. doi: 10.1177/1747021817739834

Lazaridou, A., Bruni, E., & Baroni, M. (2014). Is this a wampimuk? Cross-modal mapping between distributional semantics and the visual world. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1403–1414). Baltimore, MD: ACL.

Lazaridou, A., Marelli, M., & Baroni, M. (2017). Multimodal word meaning induction from minimal exposure to natural text. *Cognitive Science*, *41*, 677–705.

Lazaridou, A., Pham, N. T., & Baroni, M. (2015). Combining language and vision with a multimodal skip-gram model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies* (pp. 153–163). East Stroudsberg, PA.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*, 788–791.

Lemaire, B., & Denhière, G. (2004). Incremental construction of an associative network from a corpus. In *Proceedings of the Twenty-Sixth Annual Conference of the Cognitive Science Society* (pp. 825–830). Mahwah, NJ: Erlbaum.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Rivista di Linguistica*, *20*(1), 1–31.

Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics*, *4*, 151–171.

Lenci, A., & Benotto, G. (2012). Identifying hypernyms in distributional semantic spaces. In *Proceedings of *SEM* (pp. 75–79).

Lenton, A. P., Sedikides, C., & Bruder, M. (2009). A latent semantic analysis of gender stereotype-consistency and narrowness in American English. *Sex Roles*, *60*, 269–278.

Levey, A. B., & Martin, I. (1975). Classical conditioning of human 'evaluative' responses. *Behaviour Research and Therapy*, *13*, 221–226.

Levy, O., & Goldberg, Y. (2014a). Linguistic regularities in sparse and explicit word representations. In *Proceedings of the 18th Conference on Computational Natural Language Learning* (pp. 171–180). East Stroudsburg, PA: ACL.

Levy, O., & Goldberg, Y. (2014b). Neural word embedding as implicit matrix factorization. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Weinberger (Eds.), *Advances in Neural Information Processing Systems (NIPS) 2014* (Vol. 27, pp. 2177–2185).

Lewicki, P. (1986). Processing information about covariations that cannot be articulated. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*, 135–146.

Lewicki, P., Czyzewska, M., & Hoffman, H. (1987). Unconscious acquisition of complex procedural knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 523–530.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 17th International Conference on Computational Linguistics – Volume 2* (pp. 768–774). Montréal, Canada.

Ling, W., Tsvetkov, Y., Amir, S., Fermandez, R., Dyer, C., Black, A. W., … Lin, C.-C. (2015). Not all contexts are created equal: Better word representations with variable attention. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1367–1372).

Lopopolo, A., & Miltenburg, E. (2015). Sound-based distributional models. In *Proceedings of the 11th International Conference on Computational Semantics* (pp. 70–75).

Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, *3*, 273–302.

Louwerse, M. M., & Benesh, N. (2012). Representing spatial structure through maps and language: Lord of the Rings encodes the spatial structure of Middle Earth. *Cognitive Science*, *36*, 1556–1569.

Louwerse, M. M., & Connell, L. (2011). A taste of words: Linguistic context and perceptual simulation predict the modality of words. *Cognitive Science*, *35*, 381–398.

Louwerse, M. M., Raisig, S., Tillman, R., & Hutchinson, S. (2015). Time after time in words: Chronology through language statistics. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 1428–1433). Pasadena, CA.

Louwerse, M. M., & Zwaan, R. A. (2009). Language encodes geographical information. *Cognitive Science*, *33*, 51–73.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instrumentation, and Computers*, *28*, 201–208.

Lynott, D., & Ramscar, M. J. A. (2001). Can we model conceptual combination using distributional information? In *Proceedings of the Twelfth Irish Conference on Artificial Intelligence and Cognitive Science* (pp. 1–10). Maynooth, Ireland.

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.

Marelli, M., Gagné, C. L., & Spalding, T. L. (2017). Compounding as Abstract Operation in Semantic Space: A data-driven, large-scale model for relational effects in the processing of novel compounds. *Cognition*, *166*, 207–224.

Martin, D. I., & Berry, M. W. (2007). Mathematical Foundations Behind Latent Semantic Analysis. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 35–56). Mahwah, NJ: Erlbaum.

McKoon, G., & Ratcliff, R. (1998). Memory-based language processing: Psycholinguistic research in the 1990s. *Annual Review of Psychology*, *49*, 25–42.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*, 547–559.

Meteyard, L., Cuadrado, S. R., Bahrami, B., & Vigliocco, G. (2012). Coming of age: A review of embodiment and the neuroscience of semantics. *Cortex*, *48*, 788–804.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv:1301.3781v3*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems (NIPS) 2013* (Vol. 26).

Mikolov, T., Yih, W.-T., & Zweig, G. (2013). 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* (pp. 746–751). East Stroudsburg, PA: ACL.

Mitchell, J., & Lapata, M. (2010). Composition in Distributional Models of Semantics. *Cognitive Science*, *34*, 1388–1439.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, *320*, 1191–1195.

Munoz-Rubke, F., Kafadar, K., & James, K. H. (2018). A new statistical model for analyzing rating scale data pertaining to word meaning. *Psychological Research*, *82*, 787–805.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, *89*, 609–626.

Nematzadeh, A., Meylan, S. C., & Griffiths, T. L. (2017). Evaluating vector-space models of word representation, or, the unreasonable effectiveness of counting words near other words. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 859–864).

Nida, E. A. (1979). *Componential Analysis of Meaning: An In-*

troduction to Semantic Structures (2nd ed.). The Hague, NL: Mouton.

Olmos, R., Jorge-Botana, G., León, J. A., & Escudero, I. (2014). Transforming selected concepts into dimensions in latent semantic analysis. *Discourse Processes*, *51*, 494–510.

Osgood, C. E. (1952). The nature and measurement of meaning. *Psychological Bulletin*, *49*, 197–237.

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

Ouyang, L., Boroditsky, L., & Frank, M. C. (2017). Semantic coherence facilitates distributional learning. *Cognitive science*, *41*, 855–884.

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

Pantel, P., & Lin, D. (2002). Discovering word senses from text. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 613–619). Edmonton, Canada: ACM.

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543).

Pereira, F., Gershman, S., Ritter, S., & Botvinick, M. (2016). A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive Neuropsychology*, *33*, 175–190.

Perfetti, C. A. (1998). The limits of co-occurrence: Tools and theories in language research. *Discourse Processes*, *25*, 363-377.

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in Cognitive Sciences*, *10*, 233–238.

Ramscar, M., Hendrix, P., Love, B., & Baayen, R. (2013). Learning is not decline: The mental lexicon as a window into cognition across the lifespan. *The Mental Lexicon*, *8*, 450–481.

Recchia, G., & Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior Research Methods*, *41*, 647–656.

Recchia, G., & Louwerse, M. M. (2016). Archaeology through computational linguistics: inscription statistics predict excavation sites of indus valley artifacts. *Cognitive Science*, *40*, 2065–2080.

Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50). Valletta, Malta: ELRA.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.

Riordan, B., & Jones, M. N. (2011). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, *3*, 303–345.

Rogers, T. T., Lambon Ralph, M. A., Garrard, P., Bozeat, S., McClelland, J. L., Hodges, J. R., & Patterson, K. (2004). Structure and deterioration of semantic memory: a neuropsychological and computational investigation. *Psychological Review*, *111*, 205–235.

Rogers, T. T., & Wolmetz, M. (2016). Conceptual knowledge representation: A cross-section of current research. *Cognitive Neuropsychology*, *33*, 121–129.

Rommers, J., Dijkstra, T., & Bastiaansen, M. (2013). Context-dependent semantic processing in the human brain: Evidence from idiom comprehension. *Journal of Cognitive Neuroscience*, *25*, 762–776.

Roy, D., Patel, R., DeCamp, P., Kubat, R., Fleischman, M., Roy, B., ... Gorniak, P. (2006). The Human Speechtome Project. In P. Vogt, Y. Sugita, E. Tuci, & C. Nehaniv (Eds.), *Symbol Grounding and Beyond* (pp. 192–196). Berlin, Heidelberg, Germany: Springer.

Sadeghi, Z., McClelland, J. L., & Hoffman, P. (2015). You shall know an object by the company it keeps: An investigation of semantic representations derived fflamrom object co-occurrence in visual scenes. *Neuropsychologia*, *76*, 52–61.

Sadoski, M. (2018). Reading comprehension is embodied: Theoretical and practical considerations. *Educational Psychology Review*, *30*, 331–349.

Saffran, J. R. (2003). Statistical language learning mechanisms and constraints. *Current Directions in Psychological Science*, *12*, 110–114.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928.

Sahlgren, M. (2005). An introduction to random indexing. In *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering (TKE).* Copenhagen, Denmark.

Sahlgren, M. (2006). *The Word-Space-Model.* Ph.D Dissertation, Stockholm University.

Sahlgren, M. (2008). The Distributional Hypothesis. *Rivista di Linguista*, *20*, 33–53.

Schmidtke, D., Van Dyke, J. A., & Kuperman, V. (2018). Individual variability in the semantic processing of Englishcompound words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*, 421–439.

Shaoul, C., & Westbury, C. (2010). Exploring lexical co-occurrence space using HiDEx. *Behavior Research Methods*, *42*, 393–413.

Simmons, W. K., Hamann, S. B., Harenski, C. L., Hu, X. P., & Barsalou, L. W. (2008). fMRI evidence for word association and situated simulation in conceptual processing. *Journal of Physiology-Paris*, *102*, 106–119.

Sloutsky, V. M., Yim, H., Yao, X., & Dennis, S. (2017). An associative account of the development of word learning. *Cognitive Psychology*, *97*, 1–30.

Smith, E. E., & Medin, D. L. (1981). The classical view. In E. E. Smith & D. L. Medin (Eds.), *Categories and concepts* (pp. 22–60). Cambridge, MA: Harvard University Press.

Sommerauer, P., & Fokkens, A. (2018). Firearms and Tigers are Dangerous, Kitchen Knives and Zebras are Not: Testing whether Word Embeddings Can Tell. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 276–286). Brussels, Belgium: ACL.

Spranger, M. (2012). The co-evolution of basic spatial terms and categories. In L. Steels (Ed.), *Experiments in cultural language*

evolution (pp. 111–141). John Benjamins Publishing.

Spranger, M., Pauw, S., Loetzsch, M., & Steels, L. (2012). Open-ended procedural semantics. In L. Steels & M. Hild (Eds.), *Language grounding in robots* (pp. 153–172). Berlin, Heidelberg, Germany: Springer.

Steyvers, M. (2010). Combining feature norms and text data with topic models. *Acta Psychologica*, *133*, 234–243.

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, *88*, 135–170.

Tabossi, P. (1988). Accessing lexical ambiguity in different types of sentential contexts. *Journal of Memory and Language*, *27*, 324–340.

Tsvetkov, Y., Faruqui, M., Ling, W., Lample, G., & Dyer, C. (2015). Evaluation of word vector representations by subspace alignment. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 2049–2054).

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

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293–315.

Van Dam, W. O., Rueschemeyer, S.-A., & Bekkering, H. (2010). How specifically are action verbs represented in the neural motor system: an fmri study. *NeuroImage*, *53*, 1318–1325.

Van Herten, M., Chwilla, D. J., & Kolk, H. H. (2006). When heuristics clash with parsing routines: ERP evidence for conflict monitoring in sentence perception. *Journal of Cognitive Neuroscience*, *18*, 1181–1197.

Van Tiel, B., Van Miltenburg, E., Zevakhina, N., & Geurts, B. (2016). Scalar diversity. *Journal of Semantics*, *33*, 137–175.

Wade-Stein, D., & Kintsch, E. (2004). Summary street: Interactive computer support for writing. *Cognition and Instruction*, *22*, 333–362.

Westbury, C. (2016). Pay no attention to that man behind the curtain. *The Mental Lexicon*, *11*, 350–374.

Zannino, G. D., Caltagirone, C., & Carlesimo, G. A. (2015). The contribution of neurodegenerative diseases to the modelling of semantic memory: a new proposal and a review of the literature. *Neuropsychologia*, *75*, 274–290.

Zhila, A., Yih, W.-T., Meek, C., Zweig, G., & Mikolov, T. (2013). Combining heterogeneous models for measuring relational similarity. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1000–1009). East Stroudsburg, PA: ACL.

Zwaan, R. A., & Madden, C. J. (2005). Embodies sentence comprehension. In D. Pecher & R. A. Zwaan (Eds.), *Grounding cognition: The role of action and perception in memory, language, and thinking* (p. 224//245). Cambridge, UK: Cambridge University Press.