The unreasonable effectiveness of distributional semantics

Tim Lawson

University of Bristol

tim.lawson@bristol.ac.uk

1 Meaning and use

A distributional approach to semantics, which underpins virtually all computational models of language today, is frequently introduced with an appeal to Wittgenstein's remark that "the meaning of a word is its use in the language" (2009, 25^{e}). A detailed analysis of this remark and its place in Wittgenstein's philosophy is, naturally, beyond the scope of this paper. However, it will suffice to contrast it with the 'Augustinian conception of language'; specifically, that "the meaning of a word is the object it stands for" (Baker and Hacker 2005, p. 2). Traces of this picture persist in the *Tractatus* (2001), which Wittgenstein seeks to dispel in the *Investigations* (2009) by 'grammatical clarifications' of concepts such as 'meaning'. Briefly, the later Wittgenstein holds that one should think of words as tools, not names of entities, and that to understand the meaning of a word is to know how to use it, not to know the entities to which it refers (Baker and Hacker 2005, pp. 14–15). A proponent of distributional semantics might, therefore, take Wittgenstein's remark to deny a *referential* theory of meaning,[1] i.e., one in which 'meaning' is explained with reference to the world and the mind of the language user (Sahlgren 2008, p. 2).[2] Indeed, this remark prefaces the best-known expression of the *distributional hypothesis*, the theoretical principle on which the field is founded: J. R. Firth's dictum that "you shall know a word by the company it keeps" (1957, p. 11).

A more direct influence of Wittgenstein's on the development of natural language processing is less often stated. One of the pupils to whom he dictated the lectures that would become *The Blue Book* (2007) was Margaret Masterman. As Liu (2021) explains, Masterman and her colleagues in the Cambridge Language Research Unit (CLRU) pioneered machine translation and developed the first computer thesaurus (Gavin 2018, p. 653). Though these efforts prefigured the better-known origins of distributional semantics in information retrieval (Turney and Pantel 2010, pp. 143–144; Gavin 2018, pp. 655–657), if not in cognitive science (Lenci and Sahlgren 2023, pp. 14–15), they were hindered by the computational resources of the day (Liu 2021, p. 438). Furthermore, Masterman and others insisted on the importance of semantics over syntax (ibid., p. 430), which starkly opposed the dominant Chomskyan school of the 1960s (Wilks 2000, pp. 281–282; Lenci 2008, p. 5). As she put it,

[1] In relation to formal semantics (section 4), I note Wittgenstein's qualification "for a *large* class of cases of the employment of the word 'meaning' – though not for *all*" (2009, 25^{e}), which is also discussed by Baker and Hacker (2005, pp. 129–158).

[2] This paper is based on Sahlgren's doctoral thesis (2006).

"formal logic as we at present have it is not and cannot be directly relevant to the contextually-based study of semantic pattern" (Masterman 1965, IV p. 12). In Chomsky's view, however, the problem of induction necessitates an innate faculty for language (e.g. Moyal-Sharrock 2017, pp. 574–578); hence, an explanation of linguistic phenomena must be sought in terms of cognitive principles, not distributional statistics (Lenci 2008, p. 5, cf. pp. 16–17). Moreover, the idea that a formal basis could be established for ordinary language, rather than the mere analysis of its patterns, is exemplified by Richard Montague's assertion that there is "no important theoretical difference between natural languages and the artificial languages of logicians" (1974, p. 222).

The subsequent triumph of distributional methods, facilitated by the vertiginous growth in processing power and data since the 1970s, is difficult to dispute (Leiserson et al. 2020; Lenci and Sahlgren 2023, pp. 361–362). But *why* are distributional methods effective? Or, as Juan-Luis Gastaldi (2021) puts it, "Why Can Computers Understand Natural Language?" This question is the subject of the present review. As Lenci (2008, pp. 3, 14–18) explains, the distributional hypothesis has historically been subject to a 'weak' interpretation, on which distributional properties are merely correlated with latent semantic content; and a 'strong' interpretation, on which its models are explanatory theories of cognitive representations and processes (see Günther et al. 2019). This distinction has perhaps been obscured in the wake of predictive and contextual language models, whose representations are not necessarily reducible to interpretable distributional properties. While language models are sometimes understood as models of linguistic cognition (Houghton et al. 2023), I present the view that their 'unreasonable effectiveness' (Karpathy 2015) can be explained by their disclosure of the internal *structure* of language, without recourse to problematic cognitive analogies or the dissimilar phenomena of natural language acquisition and use.

I begin by outlining the foundations of distributional semantics in structural linguistics and its key concepts (section 2). Then, in section 3, I describe the origins of count-based models, and draw attention to the necessity of the segmentation of linguistic units as a precondition for distributional analysis. In section 4, I discuss compositionality and the relation between formal and distributional semantics, which contextualizes the advent of predictive language models and the continuities that obtain with their predecessors in section 5. I briefly consider contextual language models and their interpretation with respect to a structuralist view of the distributional hypothesis in section 6. Finally, in section 7, I touch on its 'strong' interpretation, evaluation methodologies, and cognitive metaphors in natural language processing. The arguments I present are predominantly based on Gastaldi (2021), Gastaldi and Pellissier (2021), Westera and Boleda (2019), and Hacker (2019); the historical and bibliographical details owe much to Sahlgren (2008), Lenci (2008), and Lenci and Sahlgren (2023), among others.

2 Distributional structure

The distributional hypothesis is typically stated in terms such as "words that occur in similar contexts tend to have similar meanings" (Turney and Pantel 2010, pp. 142–143) or, more precisely, "the degree of semantic similarity between two linguistic expressions A and B is a function of the similarity of the linguistic contexts in which A and B can appear" (Lenci 2008, p. 3). Aside from Wittgenstein and Firth (Lenci and Sahlgren 2023, pp. 10–14), the most frequent shorthand for a distributional approach

to semantics in the literature is Zellig Harris's "Distributional Structure" (1954). However, Harris was principally concerned with phonology and morphology (Lenci and Sahlgren 2023, p. 6), despite the strong association between 'distributionalism' and semantics today (Gastaldi 2021, p. 186; Gastaldi and Pellissier 2021, p. 579). This is not to say that distributionalism is not concerned with meaning, but rather that it is limited to certain of its aspects, which I revisit ahead. Rather than a general semantics, Harris's aim was to establish a methodological basis for linguistics as a science (Lenci 2008, pp. 3, 26), which "should (and, indeed, could) only deal with what is *internal* to language" (Sahlgren 2008, p. 3). As Lenci and Sahlgren explain, linguistics is unique in that "it does not have recourse to a metalanguage that is external to its object of study" (2023, p. 6). Harris's ideas are properly situated with respect to his *structuralist* intellectual heritage: he was a student of Leonard Bloomfield, the pre-eminent American structural linguist, who was deeply influenced by Ferdinand de Saussure (Matthews 2001, pp. 20–30; Gastaldi and Pellissier 2021, p. 572). We may thus look to Saussure to understand what it means for words to occur in 'similar contexts' and to have 'similar meanings'.

The term 'structural linguistics' typically refers to an intellectual trend popularized by the posthumous publication of Saussure's *Course in General Linguistics* (Matthews 2001, pp. 3–4, 6). In this work, Saussure defines the linguistic sign as an indivisible unit that relates a concept, which is *signified*, to a sound-image, which is its *signifier* (2011, pp. 65–67). Central to the theory is that the sign is 'arbitrary': there is nothing external to the language-system that determines the sound-image that stands for a given concept (ibid., pp. 67–70). Hence, a sign is distinguished solely by its functional differences from the other signs in the language system, not by the extra-linguistic relations of the concept(s) that it signifies. Such a conception of meaning is thus *differential* rather than referential (Sahlgren 2008, p. 5): like Wittgenstein, Saussure rejects the age-old view that "'words are names for 'things"' (Matthews 2001, p. 18; Lenci and Sahlgren 2023, p. 11). Saussure further separates the differences between signs into *syntagmatic* and associative or, following Hjelmslev (1938), *paradigmatic* relations. In distributional semantics, these kinds of relations between words are commonly known as first-order and second-order co-occurrence (Jurafsky and Martin 2023, p. 126), or "how typical they are as neighbors and how well they are substitutable for each other" (Schütze and Pedersen 1993, 1, cf. section 6).

The geometric definition of semantic similarity induced by vector-space models, whose representations of 'context' have been associated with syntagmatic and paradigmatic relations (Sahlgren 2008; cf. section 3), is sometimes criticized due to its breadth (Padó and Lapata 2003, p. 2). The literature thus commonly distinguishes between semantic similarity and a broader notion of word *relatedness* (Budanitsky and Hirst 2006; Jurafsky and Martin 2023, pp. 105–106); a thorough analysis of similarity measures is provided by Curran (2004). It is well-established that the neighbours of distributional representations reflect a mixture of semantic relations, e.g., synonymy and antonymy. However, as Sahlgren (2008, p. 4) notes, for this behaviour to problematize the distributional definition, one must suppose a prescriptivist view wherein these semantic relations are given *a priori* (Gastaldi 2021, p. 188). In any case, an appeal to Harris's distributionalism suggests that 'meaning', insofar as it is amenable to linguistic analysis, derives from the formal relations between signs, rather than from the material relations between referents in the world, or contextually-determined cognitive representations. In the following sections, I examine the mechanics of distributional models to assess whether and how they actualize such a conception of meaning.

3 Count-based models

Vector representations of words and concepts have a long history in machine learning and cognitive science (Turney and Pantel 2010, pp. 143–144; Lenci and Sahlgren 2023, pp. 15–16). The idea that the occurrence frequencies of words can be used to convert text into vectors is generally thought to originate in the SMART information-retrieval system (Salton 1971; cf. Lenci and Sahlgren 2023, pp. 19–20). Briefly, a system of this kind constructs vector-space representations of documents, whose dimensions are the terms of a vocabulary and whose values are the terms' frequencies. This allows a quantitative measure of the similarity of a document to a query, considered as a pseudo-document (Jurafsky and Martin 2023, pp. 108–110). In relation to the distributional hypothesis, a *vector-space model* (VSM) of this kind takes documents to be the 'similar contexts' in which words with 'similar meanings' co-occur. The insight that these vector representations may be interpreted as a term-document matrix, i.e., that its columns may be interpreted as *word vectors*, is due to Deerwester et al. (1990, pp. 394–395). Alternatively, the 'context' of a target word may be defined by a fixed-size window (Lund and Burgess 1996), or in terms of its syntactic relations (Lin 1998; Padó and Lapata 2007). Both dimensions of the resultant matrix are the terms of a vocabulary, albeit in different roles. More detailed accounts of types of context and their influences on the resulting models are given by Lenci (2018, pp. 152–154, 158) and Lenci and Sahlgren (2023, pp. 39–48).

Historically, such a model was sometimes distinguished from vector-space models by the name *distributional semantic model*, hereafter DSM (Gastaldi 2021, p. 166). Similarly, the dense vectors that result from dimensionality reduction, i.e., the transformation of co-occurrence frequencies into a lower-dimensional latent space, as well as the activations of neural networks, are typically distinguished from the sparse columns of VSMs by the name *word embeddings* (Lenci and Sahlgren 2023, p. 65; Jurafsky and Martin 2023, pp. 107, 119). Today, the term DSM is generally used to refer to any model that represents the meanings of linguistic units in terms of their distributional relations. VSMs and DSMs have been variously described as matrix factorization (Pennington et al. 2014, p. 1533), 'matrix models' (Gastaldi 2021, p. 167; Lenci and Sahlgren 2023, pp. 97–125), and *count-based* models (Baroni et al. 2014b); I will use the latter term to distinguish between models that are explicitly based on co-occurrence statistics and the predictive and contextual language models discussed in sections 5 and 6. Influential reviews of DSMs are given by Lenci (2008) and Turney and Pantel (2010), which Erk (2012) and Clark (2015) situate with respect to formal semantics (see section 4). More recently, Lenci and Sahlgren (2023) provide a textbook-length account of distributional semantics.

Sahlgren identifies term-document and term-term matrices with Saussurean syntagmatic and paradigmatic relations (2008, pp. 7–12). Gastaldi argues that this association implies a continuum between the two, i.e., that the larger a model's context window, the 'more syntagmatic' its representations (2021, p. 179). However, there is a more fundamental problem with Sahlgren's interpretation: it is contradictory to analyse the meanings of words, or any other linguistic units, in terms of their syntagmatic and paradigmatic relations, because these relations are the means by which language is divided into the units of analysis (Gastaldi 2021, pp. 196–199; Gastaldi and Pellissier 2021, p. 582).[3] Indeed, a prerequisite of virtually all the models discussed in this paper is that text is *already segmented*

[3]Notably, Gastaldi and Pellissier propose a framework in which *paradigms* are the elementary types, as opposed to grammatical types (section 4). An exposition of this work is, unfortunately, outside the purview of this paper.

into the tokens that form the models' vocabularies. Some insight into the resolution of this causal dilemma may be gleaned by returning to structural linguistics (section 2). Saussure asserts that, prior to the emergence of linguistic distinctions, 'thought' and 'sound' are a "shapeless and undifferentiated whole" (Matthews 2001, pp. 18, 78). The solution that he proposes is through the "intervention of another similarly structured albeit heterogeneous system" (Gastaldi 2021, p. 197), e.g., the concepts of our experience that we signify through language. A thread between the arguments of Gastaldi and Westera and Boleda (2019) is that this intervening system *lies beyond* the scope of distributional semantics, which I revisit in section 7.

4 Compositionality

The principle of *compositionality* in semantics is that "the meaning of an expression is a function of the meaning of its parts and of the way they are syntactically combined" (Partee 2004, p. 153). Informally, it is obvious that natural language is compositional: we can understand an expression that we have not encountered before in terms of its words and grammatical structure. In evolutionary linguistics, compositionality is a necessary condition for language acquisition and a principal object of study (e.g. Kirby 2017). Moreover, it is a fundamental feature of a formal or model-theoretic approach to semantics (Montague 1974; Partee 2004, pp. 153–181). Historically, however, compositionality has posed difficulties for distributional semantics. Intuition and generative grammar suggest that infinite well-formed expressions are possible in natural languages, but it is not possible to learn infinite (or even very many) representations. Hence, if we are to model the meanings of complex expressions, we must find a way to compose learned representations of smaller linguistic units. We perhaps also expect that the composition of representations satisfies our expectations due to the inferential aspects of language, such as entailment, which formal semantics successfully models (Lenci 2008, pp. 20–21; Boleda and Herbelot 2016, pp. 620–622). Formal and distributional semantics are thus commonly framed as complementary approaches that it would be desirable to unify (e.g. Erk 2013; Baroni et al. 2014a; Boleda 2020). However, Westera and Boleda (2019) argue that this complementarity is overstated, which I discuss after a brief overview of vector-space composition.

The simplest composition operation is vector addition. Intuitively, addition mixes the semantic content of its operands (Lenci and Sahlgren 2023, pp. 292–293). This may suffice for certain purposes, like information retrieval (Landauer and Dumais 1997, pp. 229–231), and addition remains "surprisingly good, often outperforming more sophisticated methods" (Boleda 2020, p. 10). But "language is not merely a bag of words" (Harris 1954, p. 156): commutative operations ignore word order, and plain vector operations ignore syntactic structure, both of which are obviously significant to certain aspects of meaning. The best-known study of vector-space composition operations is due to Mitchell and Lapata (2008), who compared the predictions of element-wise addition and multiplication, and weighted combinations thereof, to human judgments of similarity for pairs of intransitive verbs and their subjects. This evaluation methodology derives from Kintsch (2001)'s demonstration of a contextualization procedure for predicates (cf. section 6). In this study and its subsequent extension, the multiplication of the two vectors was most strongly correlated with human judgments (Mitchell and Lapata 2008, pp. 242–243, 2010, pp. 1414–1417), but it generalizes poorly to longer expressions (Grefenstette 2013,

p. 17). Nonetheless, addition regained popularity for the representations learned by predictive language models (section 5). Most notably, Mikolov et al. (2013b) demonstrate that the additive composition of skip-gram word embeddings produces plausible results for the parallelogram model of analogy (Rumelhart and Abrahamson 1973, cf. section 7). Mikolov et al. (2013a) went on to justify this result: if a word vector is interpreted as a representation of the probability distribution of the contexts in which it appears, then the sum of two vectors is related to the product of their context distributions, i.e., their combined probability, by the logarithmic function in the model's output layer.

Approaches to composition that go beyond a 'bag of words' by respecting word order and syntactic structure, are typically based on tensors (the generalization of vectors and matrices to orders other than 1 and 2). Related to the multiplicative approach of Mitchell and Lapata (2008, 2010) is that of Baroni and Zamparelli (2010), in which words are represented by tensors whose orders correspond to the words' grammatical types – specifically, nouns by vectors and adjectives by matrices. It is likewise restricted to expressions of a fixed grammatical structure (adjective-noun pairs). More generally, in the context of cognitive science, Smolensky (1990) proposed the non-commutative tensor or Kronecker product as a composition operation (see Smolensky and Legendre 2006). The main problems with this approach for distributional representations of words are its computational expense, due to the exponential growth in dimensionality (Grefenstette 2013, pp. 19–20), and the fact that sequences of different lengths obtain tensor representations of different orders (Clark et al. 2008, p. 1). An alternative method that avoids these problems is to learn a generic composition operation alongside the representations on which it operates and to recursively apply it to parse trees, which may thus have arbitrary length and structure (Socher et al. 2012, 2013).[4] An overview of these developments is provided by Grefenstette (2013, pp. 15–27) and Kartsaklis (2015, pp. 18–24).

Besides the dimensionality issues of the tensor product, non-commutativity alone does not account for syntactic structure. Hence, Clark and Pulman (2007, pp. 3–4) developed Smolensky's proposal to unite distributional meaning representations with a *symbolic* representation, such as a parse tree. This concept was generalized by the categorical compositional distributional framework, abbreviated as DisCoCat (Clark et al. 2008; Coecke et al. 2010). Briefly, this framework unifies grammatical type reductions with the composition of distributional representations, by exploiting the common compact closed structure of the categories of finite-dimensional vector spaces, i.e., distributional meaning representations, and Lambek's pregroup grammars (1999).[5] A more thorough account is left to Grefenstette (2013, pp. 29–44) and Kartsaklis (2015, pp. 27–38). However, the fundamental issue with these models is that either (a) the grammar must be given *a priori*, words must be annotated with grammatical types, and expressions must be well-formed; or (b) a probabilistic grammar must be learned jointly with the representations (Toumi and Koziell-Pipe 2021). In either case, the procedure is computationally expensive in comparison to traditional predictive language models (e.g. Meichanetzidis et al. 2023, p. 5). In addition, a model with a fixed grammar cannot be trivially extended to multiple languages and dialects, or phenomena like semantic change (cf. Bradley et al. 2018). Accordingly, this framework

[4] Notably, a linearized simplification of this recursive neural network architecture is equivalent to the category-theoretic approach described below (Lewis 2019).

[5] Pregroup grammars are a simplification of Lambek's original calculus (1958), which was initially regarded as a limitation of the framework. However, the authors later clarified that it is not strictly dependent on pregroup grammars, and reformulated it accordingly (Coecke et al. 2013).

awaits a large-scale or general-purpose implementation. Recent contributions to the field have leaned towards its suitability for *quantum* natural language processing, owing to the homomorphism between its string diagrams and quantum circuits (e.g. de Felice et al. 2021).

Prior to the widespread availability of pre-trained word embeddings, several authors found evidence to support a category-theoretic or tensor-based approach to composition (e.g. Baroni and Zamparelli 2010; Dinu and Lapata 2010; Grefenstette and Sadrzadeh 2011). However, these have been largely superseded by predictive language models (Milajevs et al. 2014), whose fundamental strength is that they can learn *both* syntactic and semantic relations from mostly unstructured text (cf. section 3), and thereby challenge the traditional distinction between syntax and semantics (Gastaldi 2021, pp. 186–191). Leaving aside empirical considerations, efforts to unify formal and distributional semantics are largely driven by the supposed *desideratum* of integrating the inferential and descriptive aspects of language that they respectively model. Westera and Boleda (2019) argue that this is misguided, on the basis of the distinction between 'expression' and 'speaker meaning'. In essence, formal semantics models extra-linguistic objects, i.e., the concepts of our experience that words signify, whereas distributional semantics models *linguistic* objects, the words that signify those concepts. The relations that obtain between these respective objects are naturally related, but they are not the same (see section 7). Accepting this distinction, the application of formal semantic operations to distributional representations, as in the category-theoretic approach, does not produce a general semantics. Nevertheless, distributional semantics in isolation does produce a *complete* semantics, insofar as it achieves what is possible under Harris's conception of linguistics (section 2).

5 Predictive language models

An early observation of connectionist research was that the hidden-layer activations of a neural network come to represent domain features as a by-product of its learning procedure (Rumelhart et al. 1986). The potential of these activations to serve as *distributed representations* of linguistic units was recognized long before the popularity of word embeddings (Hinton et al. 1986; Elman 1991). Generally, a combination of the per-unit input, output, and hidden-layer activations of a neural network is chosen to serve as the representations, e.g., one of the input or output layers (Bengio et al. 2006, p. 142; Mnih and Hinton 2007, p. 642; Collobert and Weston 2008, pp. 161–162), or the two are constrained to be equal (Press and Wolf 2017). A *language model* assigns probabilities to sequences of words or, equivalently, predicts upcoming words from their prior context (Jurafsky and Martin 2023, pp. 32, 134). Various statistical language models have been proposed since the 1980s, which were originally based on n-gram counts (Rosenfeld 2000). The main precursors to contemporary methods, however, are the artificial neural networks of Xu and Rudnicky (2000) and Bengio et al. (2000), which predict a target word from the words within a fixed-size context window, and the related models proposed in the 2000s (Turian et al. 2010, pp. 384–387). Despite Chomsky's famous remark that "the notion 'probability of a sentence' is an entirely useless one" (1969, p. 57), these predictive or 'neural' language models soon dominated the landscape of distributional semantics (Baroni et al. 2014b).

Initially, however, these models were computationally expensive to train and commensurately limited in the size of their vocabularies and corpora (e.g. Xu and Rudnicky 2000, p. 203; Alexandrescu

and Kirchhoff 2006, p. 1; Mnih and Hinton 2007, p. 641; Turian et al. 2010, p. 386). Their eventual dominance was heralded by the introduction of the continuous-bag-of-words and skip-gram algorithms of Mikolov et al. (2013b), generally known as `word2vec`, and the GloVe ("Global Vectors") model (Pennington et al. 2014). While it is memory-intensive to explicitly construct a co-occurrence matrix for large vocabularies and transform it into a lower-dimensional space (Turian et al. 2010, p. 385), predictive models effectively construct the lower-dimensional space *directly* and *incrementally*. This efficiency facilitated the production and dissemination of word embeddings for much larger vocabularies and corpora (Mikolov et al. 2013a, pp. 7–8; Pennington et al. 2014, p. 1536).

Despite this radical advance in the applicability of DSMs, GloVe demonstrates a degree of continuity with its count-based predecessors (section 3). Indeed, Lenci and Sahlgren (2023) categorize GloVe as a matrix model (ibid., p. 122). In the authors' view, "global matrix factorization and local context window methods" are less distinct than they appear (Pennington et al. 2014, p. 1532). Both implicitly build representations from global co-occurrence statistics, but matrix models less ably capture meaningful substructure in the embedding space, and predictive language models use global statistics less efficiently (ibid., p. 1541). In other words, "architectures play only a secondary role in [the] relation between linguistic structure and global statistics" (Gastaldi 2021, p. 164), and the contribution of predictive models is merely to optimize this encoding. A further correspondence is revealed by Levy and Goldberg (2014b), who show that the skip-gram with negative sampling algorithm (Mikolov et al. 2013b) implicitly factorizes a shifted pointwise mutual information matrix (e.g. Jurafsky and Martin 2023, pp. 116–118). The authors subsequently demonstrated that the apparent performance advantages of DSMs can be largely attributed to hyperparameter optimizations, as opposed to the inherent superiority of their architectures (Levy et al. 2015; cf. Sahlgren and Lenci 2016). Relatedly, Lenci et al. (2022) find that, contrary to the popularity of contextual models (section 6), *static* models generally outperform BERT (Devlin et al. 2019) on word-similarity and -association tasks, provided optimal hyperparameters. This conclusion is supported by Arora et al. (2020), who show that static and even *random* embeddings can achieve similar performance to contextual embeddings (see also Gupta et al. 2019, pp. 5244–5246; Bommasani et al. 2020, pp. 4760–4762).

I noted in section 2 that distributionalism is strictly agnostic with respect to the linguistic units of analysis – it does not privilege the lexical level of language (Gastaldi 2021, p. 190). Accordingly, a productive line of language-model research has been to incorporate information from levels other than the lexical. At one end of the scale, character-level recurrent neural networks date back to at least Elman (1990) and Schmidhuber and Heil (1996). Moreover, a common solution to the problem of sparsity, i.e., the 'long tail' of the word-frequency distribution, is to use sub-word units instead of words as the model's vocabulary. For instance, the `fastText` model extends the skip-gram algorithm (Mikolov et al. 2013b) to learn representations for character *n*-grams (sub-words), which are combined to form representations for words (Bojanowski et al. 2017, pp. 136–137). Similarly, the popular BERT model uses a sub-word vocabulary (Wu et al. 2016) as its input embedding layer. At the other end of the scale, supra-lexical information is intimately related to compositionality (section 4). The representation of a sequence, such as a phrase or sentence, serves as a compositional representation of its constituent items. BERT likewise associates a special `CLS` ('classification') token with a sequence, separate from its constituent sub-word tokens (Devlin et al. 2019, p. 4174).

The essential characteristic of predictive language models with respect to a distributionalist conception of meaning is that their *pre-trained* representations are able to serve as inputs or be adapted to a wide variety of tasks that involve language understanding (e.g. Turian et al. 2010; Mikolov et al. 2018; Bommasani et al. 2022, pp. 23–24). This generality supports the idea that the isolated analysis of distributional relations is sufficient to explain 'meaning' insofar as it is amenable to the science of linguistics (sections 2 and 4).

6 Contextual language models

In his *Foundations of Arithmetic*, Frege promises "never to ask for the meaning of a word in isolation, but only in the context of a proposition" (Frege 1960, p. xxii).[6] Like the principle of compositionality (section 4), this tenet is intuitive: words are frequently polysemous (e.g. Kintsch 2001, pp. 173–174), or assume different connotations and emphasis within different expressions (e.g. Armendariz et al. 2020, pp. 2–3). However, both count-based (section 3) and predictive models (section 5) originally produced a single, *static* representation for each token of the model's vocabulary, which thus encoded all of its senses and connotations (e.g. Pelevina et al. 2016, p. 1). Historically, this problem was generally addressed by one of two approaches: (a) by producing a representation for each sense of a target word and discriminating between them in the given context; or (b) by contextualizing the representation of the target word based on those of its context, without explicitly modelling word senses. In this section, I provide an overview of these approaches before discussing *contextual* language models.

The task of word-sense discrimination is to determine the sense of an ambiguous word in a given context (Schütze 1998, p. 97). Based on Miller and Charles's idea that word senses are determined by contextual similarity (cf. section 7), Schütze proposed an influential unsupervised method of word-sense discrimination (ibid., p. 99), which followed his work to model syntagmatic and paradigmatic relations (Schütze and Pedersen 1993; cf. section 2). This method clusters the contexts in which target words occur, which Reisinger and Mooney (2010) subsequently proposed to integrate with the construction of word vectors to produce a 'prototype' vector for each cluster. Kintsch (2001) proposed an analogous procedure to contextualize the meaning of a predicate according to its particular arguments, which Erk and Padó (2008) generalized by a 'structured vector-space' model that represents the selectional preferences of words with respect to arbitrary syntactic relations. A similarly syntactically-informed approach is given by Thater et al. (2010, 2011). Multi-prototype models have likewise been proposed in the predictive setting (e.g. Huang et al. 2012; Tian et al. 2014; Pelevina et al. 2016). As evidenced in section 4, these methods correspond closely to the treatment of composition in distributional semantics (e.g. Mitchell and Lapata 2008, 2009, 2010; Dinu and Lapata 2010).

However, it was quickly recognized that the activations of neural networks that apply to sequences, e.g., recurrent neural networks, could serve as *contextual* representations of the sequences' items (McCann et al. 2017; Peters et al. 2017, 2018). For these models, therefore, composition and contextualization may be relieved of their status as separate problems that necessitate methods such as the above. Today, the predominant contextual language models are based on the attention mecha-

[6] In relation to section 4, the principle of compositionality has also been attributed to Frege (Janssen 2001; Pelletier 2001). Both principles are also significant to the early and later Wittgenstein (Baker and Hacker 2005, pp. 159–188).

nism (Bahdanau et al. 2015, pp. 3–4) and, in particular, the Transformer architecture (Vaswani et al. 2017). This mechanism generalizes the ability of recurrent neural networks to represent long-range dependencies between the items of sequences. However, unlike networks with recurrent connections, the computation necessitated by Transformers is *parallelizable*, which has enabled them to be scaled to much larger corpora and many more parameters than their predecessors (e.g. Kaplan et al. 2020).

Gastaldi (2021, p. 190) notes in his argument for a 'structuralist hypothesis' that it does not treat this generation of DSMs explicitly. However, like Lenci and Sahlgren (2023, p. 355), he takes the view that they are no less founded on the theoretical principle of the distributional hypothesis than their predecessors, and we may draw an analogy between their mechanisms to support this view. In a Transformer block, each item in a sequence is represented by a *query* vector, which is compared to the *key* vectors of all the other items. The results of these comparisons are used to compute a weighted sum of the *value* vectors of the items (Vaswani et al. 2017, pp. 3–4). The analogy with the predictive models of section 5 is clear: the query vector corresponds to the target word, the key vectors to its context words, and the sum of value vectors to a contextualization procedure. Like the approach of Socher et al. (2012), this procedure is learned, rather than defined *a priori*, which provides its generality. Contextual language models likewise preserve the ability of predictive models to jointly learn syntactic and semantic information (e.g. Hewitt and Manning 2019).

In recent years, the majority of NLP research has focused on the development, application, and analysis of *foundation* models (Bommasani et al. 2022, pp. 22–27). This paradigm shift can be identified with the introduction of the GPT (Radford et al. 2018, 2019) and BERT (Devlin et al. 2019) models, which accord with the description above. While models of this kind have achieved widespread success on benchmark tasks (Bommasani et al. 2022, pp. 22–27), there is cause to criticize the suitability of typical benchmarks for characterizing their capabilities and limitations (Srivastava et al. 2023, pp. 5–6). Furthermore, the apparent emergent abilities of language models (Wei et al. 2022) on zero- and few-shot learning tasks, which are the most plausible source of their likeness to cognition, can be at least partly attributed to the increasing contamination of their training data (Li and Flanigan 2023). Hence, I discuss this likeness in the following section.

7 Cognition and evaluation

The 'strong' version of the distributional hypothesis holds that the linguistic contexts in which a word appears have a causal relationship to its cognitive representation (Lenci 2008, pp. 16–18; Lenci and Sahlgren 2023, pp. 18–19). This interpretation derives from psychology and cognitive science, where distributional methods have been widely applied (e.g. Lenci 2008, p. 16; Lenci and Sahlgren 2023, pp. 15–16; Landauer et al. 2011). Indeed, some of the most influential work in distributional semantics has been developed and evaluated in the context of cognitive science, and vector representations of meaning in psychology predate the development of vector-space models (sections 1 and 3). In response to the problem of induction, for instance, Landauer and Dumais (1997) describe Latent Semantic Analysis as a theory of how we learn and use the meanings of words (cf. Landauer et al. 2011, pp. 3–34). There is likewise a relation between the cognitive interpretation and a 'use theory of meaning' (Lenci and Sahlgren 2023, pp. 22–23). This view is notably advocated by Miller and Charles (1991), who

describes the 'contextual representation' of a word as "knowledge of how that word is used" (1991, p. 4). Hence, the 'strong' version of their 'contextual hypothesis' is that the semantic similarity of two words is determined by the similarity of their contextual, i.e., cognitive, representations (ibid., p. 8). I take the view that this interpretation of the distributional hypothesis is unwarranted, and that it is due principally to (a) the practical necessity of evaluating distributional models against human judgments and (b) the widespread usage of cognitive metaphors in the literature (Gastaldi and Pellissier 2021, pp. 569–570). I discuss these points in turn, before concluding.

The evaluation of language models may be divided into *extrinsic* evaluation, which is based on the performance of the model on a specific task, and *intrinsic* evaluation, which is notionally separate from any such task (Wang et al. 2020, p. 38; Torregrossa et al. 2021; Bommasani et al. 2022, pp. 91–96). This distinction is particularly important for the evaluation of foundation models (section 6), which are explicitly intended to be task-agnostic. Intrinsic evaluation is largely based on correspondence with human judgments, e.g., ratings of semantic similarity or relatedness (Lenci and Sahlgren 2023, p. 17), as evidenced by the survey of evaluation datasets in Lenci et al. (2022, p. 1281). For example, the word analogy task (section 4) has been used to evaluate the quality of embedding spaces (Mnih and Kavukcuoglu 2013, pp. 4–5; Levy and Goldberg 2014a; Pennington et al. 2014), but the generality of this approach has been questioned (Lenci et al. 2022, p. 1300). There is likewise cause to criticize word similarity as an evaluation methodology altogether (Batchkarov et al. 2016), not least due to the incompatibility of a naïve geometric definition with observable cognitive phenomena (Lewis 2022, pp. 3–11). Performance on an intrinsic word-similarity task does not necessarily translate to extrinsic downstream tasks (Batchkarov et al. 2016, pp. 7–8), and inter-annotator agreement is generally poor for word-similarity in comparison to more specific tasks (ibid., pp. 8–9). The correspondences between intrinsic and extrinsic evaluation are thus an important topic of research (Bommasani et al. 2022, pp. 92–94). While evaluation with respect to human judgments is practically necessary, it implies a cognitive interpretation of the model in question, which may be misleading. Accordingly, De Deyne et al. (2016) show that models explicitly based on word-association can outperform DSMs.

The language we use to describe computers and artificial-intelligence systems makes widespread use of cognitive metaphors: computers and LSTM networks have 'memory', Transformer blocks 'pay attention to' textual features, and so on. This usage largely originates in the connectionist tradition and 'neural' language models (section 5). Interpreted literally, these metaphors lead us towards the misconception that increasingly sophisticated artificial-intelligence systems are increasingly akin to the mind (Hacker 2019, p. 103). The crux of Hacker's argument against this view is that the faculties of the mind are *normative*, whereas the behaviour of machines is *causal*: following a rule is something other than executing a program, and determining whether the result of executing a program is correct is something other than verifying its 'causal inevitability' (ibid., pp. 107–108). We do not speak of less sophisticated calculating devices as 'thinking', except metaphorically, and we must take care to retain this metaphorical distinction when we speak of artificial-intelligence systems like language models. Indeed, it is difficult to see how particular DSMs could be interpreted as models of cognitive linguistic processes, since they commonly employ different architectures (Gastaldi 2021, p. 570). As Houghton et al. (2023) have noted, state-of-the-art language models are not necessarily more like the brain than their predecessors: recurrent neural networks have generally been superseded by Transformers (Wang

et al. 2019; Gillioz et al. 2020), which eschew recurrent connections for the sake of computational efficiency (section 6), but humans are generally considered to process language sequentially (Dominey et al. 2003). What is common to these models is the theoretical principle of the distributional hypothesis, and that they learn distributed representations of the 'meanings' of linguistic units (specifically, tensors in a high-dimensional vector space). Language-model training is furthermore increasingly dissimilar to natural language acquisition: rather than ingesting vast textual corpora, we use language to *interact* with each other to achieve shared goals in our environment (Lenci and Sahlgren 2023, pp. 361–363). Hence, I suggest, the successes of DSMs in modelling cognitive phenomena vindicate this theoretical principle and representational form, but not more.

The 'structuralist hypothesis' of Gastaldi (2021) and Gastaldi and Pellissier (2021) holds that distributional semantics manifests purely formal relations between linguistic units, which are strictly separate from cognitive representations and processes, and that its successes are principally due to the reality of this internal structure. We might expect that this structure bears some resemblance to that of the concepts of our experience which it signifies – perhaps even as a consequence of the cognitive processes that underlie its acquisition and use (see Goldsmith 2005). As Landauer et al. (2011, pp. 5–8) explains, language is in part a map of the world, and we can derive useful knowledge from the map alone. But "the map is not the territory" (Korzybski 1995, p. 58), and the relations between the map and the territory are perhaps beyond the scope of distributional semantics.

References

Alexandrescu, Andrei and Katrin Kirchhoff (2006). "Factored Neural Language Models". In: *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*. Ed. by Robert C. Moore et al. New York, NY: Association for Computational Linguistics, pp. 1–4.

Armendariz, Carlos Santos et al. (2020). "SemEval-2020 Task 3: Graded Word Similarity in Context". In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. Ed. by Aurélie Herbelot et al. Barcelona (online): International Committee for Computational Linguistics, pp. 36–49.

Arora, Simran et al. (2020). "Contextual Embeddings: When Are They Worth It?" In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 2650–2663.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2015). "Neural Machine Translation by Jointly Learning to Align and Translate". In: *3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. San Diego, CA.

Baker, Gordon P. and P. M. S. Hacker (2005). *Wittgenstein: Understanding and Meaning (Volume 1 of An Analytical Commentary on the Philosophical Investigations), Part 1: Essays*. 2nd ed. Vol. 1. Malden, MA: Blackwell Publishing.

Baroni, Marco, Raffaella Bernardi, and Roberto Zamparelli (2014a). "Frege in Space: A Program for Compositional Distributional Semantics". In: *Linguistic Issues in Language Technology* 9.

Baroni, Marco, Georgiana Dinu, and Germán Kruszewski (2014b). "Don't Count, Predict! A Systematic Comparison of Context-Counting vs. Context-Predicting Semantic Vectors". In: *Proceedings*

of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Baltimore, MD: Association for Computational Linguistics, pp. 238–247.

Baroni, Marco and Roberto Zamparelli (2010). "Nouns Are Vectors, Adjectives Are Matrices: Representing Adjective-Noun Constructions in Semantic Space". In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Cambridge, MA: Association for Computational Linguistics, pp. 1183–1193.

Batchkarov, Miroslav et al. (2016). "A Critique of Word Similarity as a Method for Evaluating Distributional Semantic Models". In: *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*. Berlin, Germany: Association for Computational Linguistics, pp. 7–12.

Bengio, Yoshua, Réjean Ducharme, and Pascal Vincent (2000). "A Neural Probabilistic Language Model". In: *Advances in Neural Information Processing Systems 13 (NIPS 2000)*. Ed. by Todd K. Leen, Thomas G. Dietterich, and Volker Tresp. Vol. 13. Denver, CO: Curran Associates, pp. 932–938.

Bengio, Yoshua et al. (2006). "Neural Probabilistic Language Models". In: *Innovations in Machine Learning: Theory and Applications*. Ed. by Dawn E. Holmes and Lakhmi C. Jain. 1st ed. Vol. 194. Studies in Fuzziness and Soft Computing. Heidelberg, Germany: Springer Berlin, pp. 137–186.

Bojanowski, Piotr et al. (2017). "Enriching Word Vectors with Subword Information". In: *Transactions of the Association for Computational Linguistics* 5, pp. 135–146.

Boleda, Gemma (2020). "Distributional Semantics and Linguistic Theory". In: *Annual Review of Linguistics* 6.1, pp. 213–234.

Boleda, Gemma and Aurélie Herbelot (2016). "Formal Distributional Semantics: Introduction to the Special Issue". In: *Computational Linguistics* 42.4, pp. 619–635.

Bommasani, Rishi, Kelly Davis, and Claire Cardie (2020). "Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 4758–4781.

Bommasani, Rishi et al. (2022). *On the Opportunities and Risks of Foundation Models*. URL: `http://arxiv.org/abs/2108.07258`.

Bradley, Tai-Danae et al. (2018). "Translating and Evolving: Towards a Model of Language Change in DisCoCat". In: *Electronic Proceedings in Theoretical Computer Science* 283, pp. 50–61.

Budanitsky, Alexander and Graeme Hirst (2006). "Evaluating WordNet-based Measures of Lexical Semantic Relatedness". In: *Computational Linguistics* 32.1, pp. 13–47.

Chomsky, N. (1969). "Quine's Empirical Assumptions". In: *Words and Objections: Essays on the Work of W. V. Quine*. Ed. by Donald Davidson and Jaakko Hintikka. Synthese Library. Dordrecht, The Netherlands: Springer Netherlands, pp. 53–68.

Clark, Stephen (2015). "Vector Space Models of Lexical Meaning". In: *The Handbook of Contemporary Semantic Theory*. Ed. by Shalom Lappin and Chris Fox. 2nd ed. Blackwell Handbooks in Linguistics. Chichester, UK: John Wiley & Sons, pp. 493–522.

Clark, Stephen, Bob Coecke, and Mehrnoosh Sadrzadeh (2008). "A Compositional Distributional Model of Meaning". In: *Proceedings of the Second Quantum Interaction Symposium (QI-2008)*. Ed. by Peter D. Bruza et al. Oxford, UK: College Publications, pp. 133–140.

Clark, Stephen and Stephen Pulman (2007). "Combining Symbolic and Distributional Models of Meaning". In: *Papers from the 2007 AAAI Spring Symposium*. Stanford, CA: Association for the Advancement of Artificial Intelligence, pp. 52–55.

Coecke, Bob, Edward Grefenstette, and Mehrnoosh Sadrzadeh (2013). "Lambek vs. Lambek: Functorial Vector Space Semantics and String Diagrams for Lambek Calculus". In: *Annals of Pure and Applied Logic*. Special issue on Seventh Workshop on Games for Logic and Programming Languages (GaLoP VII) 164.11, pp. 1079–1100.

Coecke, Bob, Mehrnoosh Sadrzadeh, and Stephen Clark (2010). "Mathematical Foundations for a Compositional Distributional Model of Meaning". In: *Linguistic Analysis* 36.1–4, pp. 345–384.

Collobert, Ronan and Jason Weston (2008). "A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning". In: *Proceedings of the 25th International Conference on Machine Learning - ICML '08*. Helsinki, Finland: ACM Press, pp. 160–167.

Curran, James R. (2004). "From Distributional to Semantic Similarity". PhD thesis. Edinburgh, UK: University of Edinburgh. URL: https://era.ed.ac.uk/bitstream/handle/1842/563/IP030023.pdf.

De Deyne, Simon, Amy Perfors, and Daniel J Navarro (2016). "Predicting Human Similarity Judgments With Distributional Models: The Value of Word Associations". In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Ed. by Yuji Matsumoto and Rashmi Prasad. Osaka, Japan: The COLING 2016 Organizing Committee, pp. 1861–1870.

de Felice, Giovanni, Alexis Toumi, and Bob Coecke (2021). "DisCoPy: Monoidal Categories in Python". In: *Electronic Proceedings in Theoretical Computer Science* 333, pp. 183–197.

Deerwester, Scott et al. (1990). "Indexing by Latent Semantic Analysis". In: *Journal of the American Society for Information Science* 41.6, pp. 391–407.

Devlin, Jacob et al. (2019). "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, MN: Association for Computational Linguistics, pp. 4171–4186.

Dinu, Georgiana and Mirella Lapata (2010). "Measuring Distributional Similarity in Context". In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Ed. by Hang Li and Lluís Màrquez. Cambridge, MA: Association for Computational Linguistics, pp. 1162–1172.

Dominey, Peter F. et al. (2003). "Neurological Basis of Language and Sequential Cognition: Evidence From Simulation, Aphasia, and ERP Studies". In: *Brain and Language* 86.2, pp. 207–225.

Elman, Jeffrey L. (1990). "Finding Structure in Time". In: *Cognitive Science* 14.2, pp. 179–211.

— (1991). "Distributed Representations, Simple Recurrent Networks, and Grammatical Structure". In: *Machine Learning* 7.2, pp. 195–225.

Erk, Katrin (2012). "Vector Space Models of Word Meaning and Phrase Meaning: A Survey". In: *Language and Linguistics Compass* 6.10, pp. 635–653.

— (2013). "Towards a Semantics for Distributional Representations". In: *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*. Potsdam, Germany: Association for Computational Linguistics, pp. 95–106.

Erk, Katrin and Sebastian Padó (2008). "A Structured Vector Space Model for Word Meaning in Context". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing - EMNLP '08*. Honolulu, Hawaii: Association for Computational Linguistics, p. 897.

Firth, J. R. (1957). "A Synopsis of Linguistic Theory, 1930–1955". In: *Studies in Linguistic Analysis*. Oxford, UK: Blackwell Publishing, pp. 1–32.

Frege, Gottlob (1960). *The Foundations of Arithmetic: A Logico-Mathematical Enquiry Into the Concept of Number*. Trans. by J. L. Austin. 2nd ed. New York, NY: Harper & Brothers.

Gastaldi, Juan Luis (2021). "Why Can Computers Understand Natural Language?" In: *Philosophy & Technology* 34.1, pp. 149–214.

Gastaldi, Juan Luis and Luc Pellissier (2021). "The Calculus of Language: Explicit Representation of Emergent Linguistic Structure Through Type-Theoretical Paradigms". In: *Interdisciplinary Science Reviews* 46.4, pp. 569–590.

Gavin, Michael (2018). "Vector Semantics, William Empson, and the Study of Ambiguity". In: *Critical Inquiry* 44.4, pp. 641–673.

Gillioz, Anthony et al. (2020). "Overview of the Transformer-based Models for NLP Tasks". In: *2020 Federated Conference on Computer Science and Information Systems*, pp. 179–183.

Goldsmith, John (2005). "Review of The Legacy of Zellig Harris: Language and Information into the 21st Century, vol. 1: Philosophy of Science, Syntax and Semantics". In: *Language* 81.3, pp. 719–736.

Grefenstette, Edward (2013). "Category-Theoretic Quantitative Compositional Distributional Models of Natural Language Semantics". PhD thesis. Oxford, UK: University of Oxford. URL: http://arxiv.org/abs/1311.1539.

Grefenstette, Edward and Mehrnoosh Sadrzadeh (2011). "Experimental Support for a Categorical Compositional Distributional Model of Meaning". In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Ed. by Regina Barzilay and Mark Johnson. Edinburgh, UK: Association for Computational Linguistics, pp. 1394–1404.

Günther, Fritz, Luca Rinaldi, and Marco Marelli (2019). "Vector-Space Models of Semantic Representation From a Cognitive Perspective: A Discussion of Common Misconceptions". In: *Perspectives on Psychological Science: A Journal of the Association for Psychological Science* 14.6, pp. 1006–1033.

Gupta, Prakhar, Matteo Pagliardini, and Martin Jaggi (2019). "Better Word Embeddings by Disentangling Contextual n-Gram Information". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, MN: Association for Computational Linguistics, pp. 933–939.

Hacker, P. M. S. (2019). *Wittgenstein: Meaning and Mind (Volume 3 of An Analytical Commentary on the Philosophical Investigations), Part 1: Essays*. 2nd ed. Vol. 3. Oxford, UK: Blackwell Publishing.

Harris, Zellig S. (1954). "Distributional Structure". In: *WORD* 10.2-3, pp. 146–162.

Hewitt, John and Christopher D. Manning (2019). "A Structural Probe for Finding Syntax in Word Representations". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, MN: Association for Computational Linguistics, pp. 4129–4138.

Hinton, Geoffrey E., James L. McClelland, and David E. Rumelhart (1986). "Distributed Representations". In: *Parallel Distributed Processing, Volume 1: Explorations in the Microstructure of Cognition: Foundations*. Vol. 1. Cambridge, MA: The MIT Press, pp. 77–109.

Hjelmslev, Louis (1938). "Essai d'une théorie des morphèmes". In: *Actes du Quatrième Congrès International de Linguistes*. Ed. by Ken Barr. Copenhagen, Denmark: Einar Munksgaard, pp. 140–151.

Houghton, Conor, Nina Kazanina, and Priyanka Sukumaran (2023). "Beyond the Limitations of Any Imaginable Mechanism: Large Language Models and Psycholinguistics". In: *Behavioral and Brain Sciences* 46.

Huang, Eric et al. (2012). "Improving Word Representations via Global Context and Multiple Word Prototypes". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Haizhou Li et al. Jeju Island, Korea: Association for Computational Linguistics, pp. 873–882.

Janssen, Theo M. V. (2001). "Frege, Contextuality and Compositionality". In: *Journal of Logic, Language, and Information* 10.1, pp. 115–136.

Jurafsky, Dan and James H. Martin (2023). *Speech and Language Processing*. URL: https://web.stanford.edu/~jurafsky/slp3/.

Kaplan, Jared et al. (2020). *Scaling Laws for Neural Language Models*. URL: http://arxiv.org/abs/2001.08361.

Karpathy, Andrej (2015). *The Unreasonable Effectiveness of Recurrent Neural Networks*. URL: https://karpathy.github.io/2015/05/21/rnn-effectiveness/.

Kartsaklis, Dimitri (2015). "Compositional Distributional Semantics with Compact Closed Categories and Frobenius Algebras". PhD thesis. arXiv. URL: http://arxiv.org/abs/1505.00138.

Kintsch, Walter (2001). "Predication". In: *Cognitive Science* 25.2, pp. 173–202.

Kirby, Simon (2017). "Culture and Biology in the Origins of Linguistic Structure". In: *Psychonomic Bulletin & Review* 24.1, pp. 118–137.

Korzybski, Alfred (1995). *Science and Sanity: An Introduction to Non-Aristotelian Systems and General Semantics*. 5th ed. New York, NY: Institute of General Semantics.

Lambek, Joachim (1958). "The Mathematics of Sentence Structure". In: *The American Mathematical Monthly* 65.3, pp. 154–170.

— (1999). "Type Grammar Revisited". In: *Logical Aspects of Computational Linguistics*. Ed. by Alain Lecomte, François Lamarche, and Guy Perrier. Lecture Notes in Computer Science. Berlin, Germany: Springer, pp. 1–27.

Landauer, Thomas K. and Susan T. Dumais (1997). "A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge". In: *Psychological Review* 104.2, pp. 211–240.

Landauer, Thomas K. et al., eds. (2011). *Handbook of Latent Semantic Analysis*. New York, NY: Routledge.

Leiserson, Charles E. et al. (2020). "There's Plenty of Room at the Top: What Will Drive Computer Performance After Moore's Law?" In: *Science* 368.6495.

Lenci, Alessandro (2008). "Distributional Semantics in Linguistic and Cognitive Research". In: *Rivista di Linguistica* 20.1, pp. 1–31.

— (2018). "Distributional Models of Word Meaning". In: *Annual Review of Linguistics* 4.1, pp. 151–171.

Lenci, Alessandro and Magnus Sahlgren (2023). *Distributional Semantics*. Studies in Natural Language Processing. Cambridge, UK: Cambridge University Press.

Lenci, Alessandro et al. (2022). "A Comparative Evaluation and Analysis of Three Generations of Distributional Semantic Models". In: *Language Resources and Evaluation* 56.4, pp. 1269–1313.

Levy, Omer and Yoav Goldberg (2014a). "Linguistic Regularities in Sparse and Explicit Word Representations". In: *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*. Ann Arbor, MI: Association for Computational Linguistics, pp. 171–180.

— (2014b). "Neural Word Embedding as Implicit Matrix Factorization". In: *Advances in Neural Information Processing Systems 27 (NIPS 2014)*. Ed. by Zoubin Ghahramani et al. Vol. 27. Montreal, Canada: Curran Associates, pp. 2177–2185.

Levy, Omer, Yoav Goldberg, and Ido Dagan (2015). "Improving Distributional Similarity with Lessons Learned from Word Embeddings". In: *Transactions of the Association for Computational Linguistics* 3, pp. 211–225.

Lewis, Martha (2019). "Compositionality for Recursive Neural Networks". In: *Journal of Applied Logics* 6.4, pp. 709–724.

— (2022). "Quantum Computing and Cognitive Simulation". In: *Quantum Computing in the Arts and Humanities: An Introduction to Core Concepts, Theory and Applications*. Ed. by Eduardo Reck Miranda. Cham, Switzerland: Springer International Publishing, pp. 53–105.

Li, Changmao and Jeffrey Flanigan (2023). *Task Contamination: Language Models May Not Be Few-Shot Anymore*. URL: http://arxiv.org/abs/2312.16337.

Lin, Dekang (1998). "Automatic Retrieval and Clustering of Similar Words". In: *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*. Montreal, Canada: Association for Computational Linguistics, pp. 768–774.

Liu, Lydia H. (2021). "Wittgenstein in the Machine". In: *Critical Inquiry* 47.3, pp. 425–455.

Lund, Kevin and Curt Burgess (1996). "Producing High-Dimensional Semantic Spaces from Lexical Co-occurrence". In: *Behavior Research Methods, Instruments, & Computers* 28.2, pp. 203–208.

Masterman, Margaret (1965). "Semantic Algorithms". In: *Proceedings of the Conference on Computer-Related Semantics*. Vol. 4. Las Vegas, NV, pp. 1–97.

Matthews, P. H. (2001). *A Short History of Structural Linguistics*. Cambridge, UK: Cambridge University Press.

McCann, Bryan et al. (2017). "Learned in Translation: Contextualized Word Vectors". In: *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. Ed. by Isabelle Guyon et al. Vol. 30. Long Beach, CA: Curran Associates, pp. 6294–6305.

Meichanetzidis, Konstantinos et al. (2023). "Grammar-Aware Sentence Classification on Quantum Computers". In: *Quantum Machine Intelligence* 5.1, p. 10.

Mikolov, Tomáš et al. (2013a). "Distributed Representations of Words and Phrases and Their Compositionality". In: *Advances in Neural Information Processing Systems 26 (NIPS 2013)*. Ed. by Christopher J. C. Burges et al. Vol. 26. Lake Tahoe, NV: Curran Associates, pp. 3111–3119.

Mikolov, Tomáš et al. (2013b). "Efficient Estimation of Word Representations in Vector Space". In: *1st International Conference on Learning Representations, ICLR 2013, Workshop Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. Scottsdale, AZ.

Mikolov, Tomáš et al. (2018). "Advances in Pre-Training Distributed Word Representations". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA).

Milajevs, Dmitrijs et al. (2014). "Evaluating Neural Word Representations in Tensor-Based Compositional Settings". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Alessandro Moschitti, Bo Pang, and Walter Daelemans. Doha, Qatar: Association for Computational Linguistics, pp. 708–719.

Miller, George A. and Walter G. Charles (1991). "Contextual Correlates of Semantic Similarity". In: *Language and Cognitive Processes* 6.1, pp. 1–28.

Mitchell, Jeff and Mirella Lapata (2008). "Vector-Based Models of Semantic Composition". In: *Proceedings of ACL-08: HLT*. Ed. by Johanna D. Moore et al. Columbus, OH: Association for Computational Linguistics, pp. 236–244.

— (2009). "Language Models Based on Semantic Composition". In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, pp. 430–439.

— (2010). "Composition in Distributional Models of Semantics". In: *Cognitive Science* 34.8, pp. 1388–1429.

Mnih, Andriy and Geoffrey Hinton (2007). "Three New Graphical Models for Statistical Language Modelling". In: *Proceedings of the 24th International Conference on Machine Learning*. ICML '07. New York, NY: Association for Computing Machinery, pp. 641–648.

Mnih, Andriy and Koray Kavukcuoglu (2013). "Learning Word Embeddings Efficiently with Noise-Contrastive Estimation". In: *Advances in Neural Information Processing Systems 26 (NIPS 2013)*. Ed. by Christopher J. C. Burges et al. Vol. 26. Lake Tahoe, NV: Curran Associates, pp. 2265–2273.

Montague, Richard (1974). *Formal Philosophy: Selected Papers of Richard Montague*. New Haven, CT: Yale University Press.

Moyal-Sharrock, Danièle (2017). "Universal Grammar: Wittgenstein Versus Chomsky". In: *A Companion to Wittgenstein on Education: Pedagogical Investigations*. Ed. by Michael A. Peters and Jeff Stickney. Singapore: Springer Singapore, pp. 573–599.

Padó, Sebastian and Mirella Lapata (2003). "Constructing Semantic Space Models from Parsed Corpora". In: *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. Sapporo, Japan: Association for Computational Linguistics, pp. 128–135.

— (2007). "Dependency-Based Construction of Semantic Space Models". In: *Computational Linguistics* 33.2, pp. 161–199.

Partee, Barbara H. (2004). *Compositionality in Formal Semantics: Selected Papers by Barbara H. Partee*. Ed. by Susan Rothstein. Explorations in Semantics. Malden, MA: Blackwell Publishing.

Pelevina, Maria et al. (2016). "Making Sense of Word Embeddings". In: *Proceedings of the 1st Workshop on Representation Learning for NLP*. Ed. by Phil Blunsom et al. Berlin, Germany: Association for Computational Linguistics, pp. 174–183.

Pelletier, Francis Jeffry (2001). "Did Frege Believe Frege's Principle?" In: *Journal of Logic, Language, and Information* 10.1, pp. 87–114.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning (2014). "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Alessandro Moschitti, Bo Pang, and Walter Daelemans. Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543.

Peters, Matthew E. et al. (2017). "Semi-Supervised Sequence Tagging with Bidirectional Language Models". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Regina Barzilay and Min-Yen Kan. Vancouver, Canada: Association for Computational Linguistics, pp. 1756–1765.

Peters, Matthew E. et al. (2018). "Dissecting Contextual Word Embeddings: Architecture and Representation". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Ed. by Ellen Riloff et al. Brussels, Belgium: Association for Computational Linguistics, pp. 1499–1509.

Press, Ofir and Lior Wolf (2017). "Using the Output Embedding to Improve Language Models". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Ed. by Mirella Lapata, Phil Blunsom, and Alexander Koller. Valencia, Spain: Association for Computational Linguistics, pp. 157–163.

Radford, Alec et al. (2018). *Improving Language Understanding by Generative Pre-Training*. URL: `https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf`.

Radford, Alec et al. (2019). *Language Models are Unsupervised Multitask Learners*. URL: `https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf`.

Reisinger, Joseph and Raymond J. Mooney (2010). "Multi-Prototype Vector-Space Models of Word Meaning". In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Ed. by Ron Kaplan et al. Los Angeles, CA: Association for Computational Linguistics, pp. 109–117.

Rosenfeld, R. (2000). "Two Decades of Statistical Language Modeling: Where Do We Go from Here?" In: *Proceedings of the IEEE* 88.8, pp. 1270–1278.

Rumelhart, David E. and Adele A. Abrahamson (1973). "A Model for Analogical Reasoning". In: *Cognitive Psychology* 5.1, pp. 1–28.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams (1986). "Learning Representations by Back-Propagating Errors". In: *Nature* 323, pp. 533–536.

Sahlgren, Magnus (2006). *The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations Between Words in High-Dimensional Vector Spaces*. SICS Dissertation Series 44. Stockholm, Sweden: Department of Linguistics, Stockholm University.

— (2008). "The Distributional Hypothesis". In: *Italian Journal of Disability Studies* 20, pp. 33–53.

Sahlgren, Magnus and Alessandro Lenci (2016). "The Effects of Data Size and Frequency Range on Distributional Semantic Models". In: *Proceedings of the 2016 Conference on Empirical Methods in Nat-*

ural Language Processing. Ed. by Jian Su, Kevin Duh, and Xavier Carreras. Austin, TX: Association for Computational Linguistics, pp. 975–980.

Salton, Gerard (1971). *The SMART Retrieval System: Experiments in Automatic Document Processing*. Englewood Cliffs, NJ: Prentice-Hall.

Saussure, Ferdinand de (2011). *Course in General Linguistics*. Ed. by Perry Meisel. Trans. by Wade Baskin Edited by Perry Meisel and Haun Saussy. New York, NY: Columbia University Press.

Schmidhuber, J. and S. Heil (1996). "Sequential Neural Text Compression". In: *IEEE Transactions on Neural Networks* 7.1, pp. 142–146.

Schütze, Hinrich (1998). "Automatic Word Sense Discrimination". In: *Computational Linguistics* 24.1. Ed. by Julia Hirschberg, pp. 97–123.

Schütze, Hinrich and Jan Pedersen (1993). "A Vector Model for Syntagmatic and Paradigmatic Relatedness". In: *Proceedings of the 9th Annual Conference of the UW Centre for the New OED and Text Research*, pp. 104–113.

Smolensky, Paul (1990). "Tensor Product Variable Binding and the Representation of Symbolic Structures in Connectionist Systems". In: *Artificial Intelligence* 46.1-2, pp. 159–216.

Smolensky, Paul and Géraldine Legendre (2006). *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar Volume I: Cognitive Architecture*. Vol. 1. Cambridge, MA: The MIT Press.

Socher, Richard et al. (2012). "Semantic Compositionality through Recursive Matrix-Vector Spaces". In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Jeju Island, Korea: Association for Computational Linguistics, pp. 1201–1211.

Socher, Richard et al. (2013). "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, WA: Association for Computational Linguistics, pp. 1631–1642.

Srivastava, Aarohi et al. (2023). "Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models". In: *Transactions on Machine Learning Research*.

Thater, Stefan, Hagen Fürstenau, and Manfred Pinkal (2010). "Contextualizing Semantic Representations Using Syntactically Enriched Vector Models". In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: Association for Computational Linguistics, pp. 948–957.

— (2011). "Word Meaning in Context: A Simple and Effective Vector Model". In: *Proceedings of 5th International Joint Conference on Natural Language Processing*. Chiang Mai, Thailand: Asian Federation of Natural Language Processing, pp. 1134–1143.

Tian, Fei et al. (2014). "A Probabilistic Model for Learning Multi-Prototype Word Embeddings". In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Ed. by Junichi Tsujii and Jan Hajic. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, pp. 151–160.

Torregrossa, François et al. (2021). "A Survey on Training and Evaluation of Word Embeddings". In: *International Journal of Data Science and Analytics* 11.2, pp. 85–103.

Toumi, Alexis and Alex Koziell-Pipe (2021). *Functorial Language Models*. URL: http://arxiv.org/abs/2103.14411.

Turian, Joseph, Lev-Arie Ratinov, and Yoshua Bengio (2010). "Word Representations: A Simple and General Method for Semi-Supervised Learning". In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Ed. by Jan Hajič et al. Uppsala, Sweden: Association for Computational Linguistics, pp. 384–394.

Turney, Peter D. and Patrick Pantel (2010). "From Frequency to Meaning: Vector Space Models of Semantics". In: *Journal of Artificial Intelligence Research* 37.1, pp. 141–188.

Vaswani, Ashish et al. (2017). "Attention Is All You Need". In: *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. Ed. by Isabelle Guyon et al. Vol. 30. Long Beach, CA: Curran Associates, pp. 5998–6008.

Wang, Chenguang, Mu Li, and Alexander J. Smola (2019). *Language Models with Transformers*. URL: http://arxiv.org/abs/1904.09408.

Wang, Shirui, Wenan Zhou, and Chao Jiang (2020). "A Survey of Word Embeddings Based on Deep Learning". In: *Computing* 102.3, pp. 717–740.

Wei, Jason et al. (2022). *Emergent Abilities of Large Language Models*. URL: http://arxiv.org/abs/2206.07682.

Westera, Matthijs and Gemma Boleda (2019). "Don't Blame Distributional Semantics if it can't do Entailment". In: *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*. Ed. by Simon Dobnik, Stergios Chatzikyriakidis, and Vera Demberg. Gothenburg, Sweden: Association for Computational Linguistics, pp. 120–133.

Wilks, Yorick (2000). "Margaret Masterman". In: *Early Years in Machine Translation: Memoirs and Biographies of Pioneers*. Ed. by John W. Hutchins. Studies in the History of the Language Sciences. Amsterdam, The Netherlands: John Benjamins Publishing Company, pp. 279–297.

Wittgenstein, Ludwig (2001). *Tractatus Logico-Philosophicus*. Trans. by David F. Pears and Brian F. McGuinness. London, UK: Routledge Classics.

— (2007). *Preliminary Studies for the "Philosophical Investigations": Generally known as The Blue and Brown Books*. 2nd ed. Malden, MA: Blackwell Publishing.

— (2009). *Philosophical Investigations*. Ed. by P. M. S. Hacker and Joachim Schulte. 4th ed. Chichester, UK: Wiley-Blackwell.

Wu, Yonghui et al. (2016). *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*. URL: http://arxiv.org/abs/1609.08144.

Xu, Wei and Alex Rudnicky (2000). "Can Artificial Neural Networks Learn Language Models?" In: *6th International Conference on Spoken Language Processing (ICSLP 2000)*. Vol. 1. Beijing, China: ISCA, pp. 202–205.