



ALESSANDRO LENCI AND MAGNUS SAHLGREN. Distributional Semantics (Studies in Natural Language Processing) (Hard cover). Cambridge: Cambridge University Press, 2023. 452 pp.

James Pustejovsky

To cite this article: James Pustejovsky (2024) ALESSANDRO LENCI AND MAGNUS SAHLGREN. Distributional Semantics (Studies in Natural Language Processing) (Hard cover). Cambridge: Cambridge University Press, 2023. 452 pp., WORD, 70:4, 291-310, DOI: [10.1080/00437956.2024.2414543](https://doi.org/10.1080/00437956.2024.2414543)

To link to this article: <https://doi.org/10.1080/00437956.2024.2414543>



© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 17 Dec 2024.



[Submit your article to this journal](#)



Article views: 1005



[View related articles](#)



[View Crossmark data](#)

James Pustejovsky *

Computer Science, Department of Computer Science, Brandeis University, Waltham, MA, 02453, USA

ALESSANDRO LENCI and MAGNUS SAHLGREN. *Distributional Semantics (Studies in Natural Language Processing)* (Hard cover). Cambridge: Cambridge University Press, 2023. 452 pp.

1. Introduction

It has been 70 years since the publication in this journal of Harris’s article “Distributional Structure”, a work which has become one of the foundational documents for contemporary computational linguistics. The present article reviews what could be viewed as Harris’s legacy, as presented in *Distributional Semantics* by Alessandro Lenci and Magnus Sahlgren (henceforth LS). That work is a comprehensive and compelling exploration of the theoretical and computational underpinnings of distributional semantics. It is a significant contribution to both linguistic theory and computational linguistics in its own right, providing a bridge between semantic theories and practical applications in natural language processing (NLP). As both a linguist and computational linguist, I found this book relevant to current computational as well as formal semantic research, which increasingly employs distributional methods to understand and model language meaning.

Distributional semantics relies on the principle that words that occur in similar contexts tend to have similar meanings, a technique that has become a cornerstone of modern NLP. This book addresses the growing need for a comprehensive understanding of this paradigm, offering insights that are crucial for both theoretical developments and practical applications. My review aims to highlight the importance of this book for ongoing research in both linguistics and computational linguistics, examining its major sections and topics while providing a critical analysis of its contributions in the context of linguistic research over the past 70 years. I begin by making some general observations regarding the book’s aim and scope, followed by a detailed review and examination of the themes covered in each of the three major sections of the book. I emphasize throughout how the topics and their treatment are related and relevant to contemporary research in linguistics and computational linguistics. I conclude with some general observations on the nature and use of data in linguistics.

*Email: pustejovsky@gmail.com; jamesp@cs.brandeis.edu

2. General remarks and overview

2.1. *Part 1 – Foundational theories*

Part 1 begins by laying the groundwork for distributional semantics, discussing the distributional hypothesis and its historical context. LS review the philosophical and cognitive underpinnings of distributional methods, explaining how meaning can be derived from usage patterns in language. This section covers foundational theories such as Harris’ original distributional hypothesis and Firth’s contextual theory of meaning (Firth 1951; Harris 1951, 1954). The authors discuss how these ideas have evolved and been refined over time, leading to the current understanding within the field of the term “distributional semantics”.

In addition to reviewing the structural foundations for distributional linguistic research, they also discuss the cognitive basis for distributional models, drawing on evidence from psycholinguistics and cognitive science. This interdisciplinary approach underscores the relevance of distributional semantics beyond purely computational contexts, highlighting its applicability to understanding human language processing. This becomes relevant for better appreciating the theoretical motivations behind the computational models presented later in the book.

2.2. *Part 2 – Models*

The authors provide a detailed overview of the major computational techniques used in distributional semantics. They cover vector space models, including the construction and manipulation of distributional vectors. This section also addresses key algorithms and methods such as Singular Value Decomposition (SVD), neural embeddings, and context-sensitive models. The discussion on neural embeddings, including word2vec, GloVe, ELMo, and BERT, is particularly thorough, offering insights into the mechanisms that enable these models to capture semantic distinctions in language use.

LS review how to construct and interpret distributional vectors, detailing the mathematical foundations and practical considerations involved. The authors also address the challenges associated with high-dimensional spaces and the techniques used to mitigate these issues, such as dimensionality reduction and regularization. This discussion is useful as a general resource for implementing or refining distributional models, providing both the theoretical background and practical guidance needed to use these methods.

2.3. *Part 3 – Practice*

LS demonstrate how distributional semantics can be applied to various NLP tasks, including word sense disambiguation, semantic similarity, and machine translation. They provide linguistic and NLP examples that illustrate the practical utility of distributional models. The authors address the integration of distributional semantics with other NLP techniques, such as syntax-based methods and deep learning approaches, while reflecting on how distributional models can be enhanced by incorporating additional linguistic information. This section includes examples that help highlight the utility and application of distributional semantic techniques. The examples are drawn from a broad range of linguistic phenomena and computational tasks, and help demonstrate specific aspects of the theory, such as its ability to capture

polysemy and synonymy, or its effectiveness in modeling diverse phenomena over large-scale text corpora. They also explore the use of distributional semantics in multilingual contexts, where these models are applied to languages other than English. The multilingual application of NLP techniques has grown increasingly important, while the development and training of Large Language Models (LLMs) for low-resource languages is of huge concern, which we return to later in this review. Generally, the inclusion of practical examples helps to ground the theoretical discussions, making the book both informative and relevant.

2.4. *Assessment*

Distributional Semantics is a work that is both suitable as an introduction to the field for the beginner, while also functioning as a comprehensive resource for both traditional and contemporary techniques in analyzing and encoding distributional representations of language. The book's interdisciplinary approach, drawing on insights from linguistics, cognitive science, and computer science, is another significant strength. The comprehensive coverage of both traditional vector space models and contemporary neural embeddings ensures that readers encounter a broad range of distributional techniques. Their clear explanations of mathematical concepts and algorithms, accompanied by practical examples, make the book a valuable resource for understanding distributional semantics. The inclusion of both foundational and current references as well as discussions of recent advances further enhances the book's relevance and utility.

3. Review of specifics and content

3.1. *The foundations of distributional semantics*

Chapter One opens with an introduction to distributional semantics as a field where semantic generalizations are based on the distribution of linguistic elements within language. The core principle, known as the “Distributional Hypothesis”, suggests that lexemes (words or phrases) with similar linguistic contexts share similar meanings. This hypothesis forms the backbone of distributional semantics, which has evolved into a significant theoretical and computational framework within linguistics. The roots of distributional semantics can be traced back to the work of American structuralists, particularly Harris. Harris pioneered the distributional methodology, which posits that the fundamental elements of language can be identified based on their distribution – how they occur relative to other elements in linguistic contexts. His work, especially “Methods in Structural Linguistics” (1951) and “Distributional Structure” (1954), emphasized the empirical and scientific basis of linguistics, advocating for a descriptive approach that avoids reliance on mentalistic explanations of meaning.

It is important to make a distinction between the application of distributional methods for grammatical induction or the discovery of syntactic constructions and how distributional techniques are currently used to make claims about natural language semantics. As the authors note, already by the 1940s, differentiation in terms of syntagmatic and paradigmatic distinctions was a technique being used by European Structuralist, such as Hjelmslev and Jakobson, and similar distributional methods were being developed within American structuralist circles, not only for

phonemics and morphophonemics, but also for syntactic analysis at the sentence level (Bloomfield 1933). Harris (1946), Wells (1947) and others further refined these techniques to explore the structure of sentences through immediate constituent analysis (ICA) and substitutability, in effect, merging syntagmatic and paradigmatic forms. The idea was that, by observing patterns of word co-occurrences, linguists could make inferences about syntactic categories and groupings within a sentence (e.g., subject, verb, object) based on distributional properties. In this developing framework, the focus was on formal properties of language and grouping words into syntactic categories based on distribution rather than their meaning. Both Wells and Harris were arriving at similar methods of constituent analysis through the substitution of terms in lexical sequences (Harris 1946; Wells 1947). As Harris puts it: the sole operation is “substitution, repeated time and again” (1946, quoted in Matthews 1993). Hence, while the distributional patterns are more closely tied to form than to content, Harris believed that the distribution of forms reflected deeper semantic relationships: “Difference of meaning correlates with difference of distribution (Harris 1954).”

Approaching word distributions from a very different tradition, Firth (1951, [1957] 1968) viewed word “context” to include not only immediate neighboring words, but also broader situational and cultural factors. His *contextual theory of meaning* could be summarized as “meaning = function in context”. At the surface form of utterances, this amounts to the notion that the meaning of a word can be understood by analyzing the habitual co-occurrence of the words it appears with in text, i.e., its *collocations*. Firth’s famous phrase, “You shall know a word by the company it keeps,” underpins this concept of collocations, which has become integral to distributional semantics.

Firth’s methods were later greatly expanded upon by John Sinclair, who introduced computational methods to extract and analyze collocations, thereby laying the groundwork for quantitative approaches in distributional semantics. The physical legacy of Sinclair’s interpretation of distributional semantics can be seen in the COBUILD Project (Moon 2007; Sinclair 1991), an approach to creating dictionaries based on actual usage rather than prescriptive norms. Central to his theory was an enrichment to Firth’s notion of collocation, together with a novel notion of “idioms” as fixed or semi-fixed multiword expressions, occurring together in predictable patterns. This insight has had profound consequences for NLP, and the interpretation of multiword expressions (MWE) continues to be a difficult semantic interpretive problem in Computational Linguistics (CL) (Constant et al. 2017; Milić and Schulte im Walde 2024).

A related thread not directly discussed by LS is Hanks’ subsequent development of Firth’s and Sinclair’s contextual meaning approach to distributional semantics (Hanks 2004, 2013; Jezek and Hanks 2010). As Sinclair’s collaborator on COBUILD, Hanks explored meaning as distinguished between conventional usage patterns of words (norms) and how speakers can creatively deviate from these patterns (exploitations). His work on Corpus Pattern Analysis (CPA) involved cataloging the syntagmatic patterns that words participate in, and then associating these with specific word senses (Hanks and Pustejovsky 2005; Pustejovsky, Hanks, and Rumshisky 2004), and subsequent compilation into a “patterns dictionary of English verbs” (Hanks 2004).

In this context, Wittgenstein is briefly mentioned, since his later philosophical works promote language as use and interaction. While the usage-based perspective of language-games does align with the Distributional Hypothesis, there is little evidence Wittgenstein ever framed the determination of meaning in actual distributional terms. On the other hand, Firth ([1957] 1968) repeatedly invokes Wittgenstein in his works, making explicit connections between language as use in situated contexts: “The sentence is composed of the words and that is enough (Wittgenstein 1953)”.

LS discuss the application of distributional analysis in psychology, particularly in the study of word associations. Early psychological experiments used word association tests to measure the strength of association between words, revealing that words with similar distributions in language tend to be associated with one another. This was influenced by and aligns the distributional approach to language analysis, where semantic similarity is seen as a function of distributional similarity. Osgood’s work on semantic space is mentioned as a precursor to vector-based representations in distributional semantics (Osgood, Suci, and Tannenbaum 1957). Osgood’s method, known as the semantic differential, involved asking subjects to rate words along various bipolar scales (e.g., happy-sad) to map their positions in a multidimensional space. While not purely distributional, this approach laid the groundwork for later computational models that represent words as vectors in a semantic space.

Having presented the origin story and its subsequent development, the authors distinguish two primary interpretations of the Distributional Hypothesis, *weak* and *strong*. Under the former interpretation, the distribution of words in contexts *correlates* with their semantic properties. It serves as an empirical methodology for analyzing lexical meaning, where the focus is on identifying patterns in linguistic data that reflect underlying semantic similarities.

Under the latter perspective, linguistic distributions are interpreted as actively *shaping* and *determining* the cognitive representations of meaning in the brain. According to this view, distributional semantics is not just a tool for semantic analysis but also for understanding how meaning is acquired and represented in the mental lexicon. Which perspective you take has obvious consequences for your goals when interpreting the results of distributional semantics.

Chapter Two lays out the principles of distributional semantic models. LS define the core of distributional semantics as the construction of distributional vectors that represent the lexemes of a language in high-dimensional spaces. These vectors are representations of the co-occurrence data of a target word and its “context”, as found in (or compiled from) a corpus. The result of such a computation is a co-occurrence matrix, where each row represents a target word, and each column is the context being examined for this word. In computational linguistics research, contexts are the co-occurring words in some window (e.g., n tokens; the sentence); this contrasts with early information retrieval work, where contexts could be document-level values. The goal is to measure how often lexemes co-occur with certain contexts. Contexts can be syntactic (relations like subject-object dependencies) or simply based on proximity (window-based methods).

Since not all collocations of a word contribute equally to determine its contextual meaning, some sort of “association metric” is usually employed to determine the importance of a particular co-occurrence characterizing the target word. For example, Pointwise Mutual Information (PMI), introduced by Church and Hanks

(1989), captures how surprising a co-occurrence is, given the overall distributions of the target and the context. As LS point out, weighting functions are critical for mitigating the skew introduced by very frequent words (like stop words) that add little to determining the meaning of a word. In addition to association measures, “relevance measures” such as *tf-idf* (Term Frequency-Inverse Document Frequency) can help gauge the importance of a word to a document, relative to a collection (or corpus) of documents.

LS introduce the fundamental notion that the co-occurrence matrix of a lexis (e.g., the set of unique words in a corpus) represents words’ distributional vectors in a vector space. Without any “pruning”, the context size for words – and hence the dimensionality of the vector space – in a language is proportional to its vocabulary, making explicit vectors (Levy and Goldberg 2014) high-dimensional; they are also sparse. Hence, there are formal and computational motivations for reducing the dimensionality of word vectors, to produce low-dimensional embeddings that capture the “latent semantic structure” associated with a word. One of the most fundamental dimensionality reduction techniques is *Singular Value Decomposition* (SVD), a matrix factorization strategy used to reduce the dimensionality by decomposing the co-occurrence matrix into factored matrices, reducing the complexity of data by representing it in terms of its most important patterns, making it easier to analyze and work with. This results in what LS identify as a *word embedding*, a low-dimensional, implicit distributional vector. The concept is based on mapping the rows of a co-occurrence matrix (which initially represent explicit high-dimensional distributional vectors) onto a reduced matrix, where the columns represent latent dimensions. These latent dimensions capture deeper semantic relationships between words that are not directly observable in the explicit distributional vector.

Once distributional vectors have been constructed for the lexis of a corpus, some means of comparing them to each is required. The most common similarity metric is *cosine similarity*, which measures the cosine of the angle between two vectors. This is popular and in fact preferred because it normalizes the vector magnitudes and focuses on the direction, which reflects the distributional similarity of the lexemes rather than their frequency. Another measure used is the Euclidean distance between vectors, but this is less common in distributional semantics due to its sensitivity to vector magnitude, which correlates with lexeme frequency.

This chapter provides a thorough examination of the foundations of distributional semantics by discussing the algorithms and techniques used to extract, represent, and compare distributional vectors. Formal measures such as PMI, vector similarity as measured by cosine values, and SVD are key tools in building effective and meaningful representations of lexical meaning. However, as the authors note, the trade-offs between computational complexity, data sparsity, and interpretability remain central challenges in this area of research.

3.2. *Distributional semantic models (DSMs)*

In Part 2, LS introduce Distributional Semantic Models (DSMs) as computational models that learn distributional representations of target lexemes by encoding their co-occurrence statistics with linguistic contexts. DSMs are defined in terms of two dimensions: the type of linguistic context assumed, and the method employed to learn distributional vectors. Context refers to the distinction encountered earlier

between: *linguistically-defined units* (words, phrases) along with further distinctions based on window-based models, syntax-based models, and construction models; and *document models* which assume documents or larger discourse units as contexts, which can lead to significant differences in the resulting semantic spaces.

Chapter Four introduces the classic matrix model for DSMs and introduces the two major methods that have been developed for the construction of distributional semantic models: (a) *count models*, which record and count co-occurrences to build distributional vectors. The classic example is the Vector Space Model (VSM); and (b) *predict models*, which use neural networks to generate dense vectors (embeddings) by predicting contexts, obviating the need to construct large co-occurrence matrices. A major difference within predict models stems from how embeddings are constructed, resulting in a distinction between static vs. contextual embeddings. Static embeddings represent words out-of-context (e.g., Word2vec, FastText), while contextual embeddings (e.g., BERT, GPT, RoBERTa) are adjusted and sensitive to specific linguistic contexts.

Matrix models are foundational in DSMs, developed from VSMs in information retrieval. These models rely on a co-occurrence matrix where each element represents the frequency of a target lexeme in a specific context. The standard model is a classical DSM that represents words as vectors derived from a window-based co-occurrence matrix, using Pointwise PMI, and vector similarity measured with cosine similarity.

Some of the key matrix models covered include: (a) Hyperspace Analogue to Language (HAL), a window-based DSM that uses raw co-occurrence counts and is weighted by a linear ramping function based on distance from the target word. Words that are farther from the target receive progressively lower weights (Lund and Burgess 1996); Latent Semantic Analysis (LSA), which uses Singular Value Decomposition (SVD) to reduce the dimensionality of the co-occurrence matrix, a process which captures latent semantic relationships between words by abstracting away from surface-level co-occurrences and finding the deeper structure in the data (Landauer and Dumais 1997).

In the context of count models, LS discuss the hybrid model, GloVe (Global Vectors for Word Representation), because it relies heavily on co-occurrence statistics from a corpus. It explicitly captures these global statistics, hence the “Global” in its name. However, it is not strictly a count model in the same way as HAL or LSA, since it incorporates these co-occurrence statistics (the counts) into a predictive (task) framework through matrix factorization, compressing them into dense vectors.

Chapter Five explores Random Encoding Models, which aim to efficiently create low-dimensional word embeddings by leveraging random projections and indexing. Unlike traditional matrix models that first construct a high-dimensional co-occurrence matrix, random encoding models skip this step, directly generating dense vectors that emphasize scalability and incrementality.

The technical methods discussed include an approach providing “dimensionality reduction with low distortion”, known as the Johnson-Lindenstrauss lemma (Johnson and Lindenstrauss 1984). This states that it is possible to embed a set of high-dimensional points into a much lower-dimensional space, in a way that approximately preserves the distances between the points. This provides a kind of formal guarantee for the vector representations constructed by Random Projection techniques. One drawback is that it still requires constructing an initial co-occurrence

matrix, which can be potentially computationally expensive when scaling to large corpora.

Random Indexing (RI) sidesteps the need for a full co-occurrence matrix, while sharing with Random Projection the properties of high-dimensional spaces, such as the tendency of random vectors to be nearly orthogonal (Kanerva 1988). RI offers a scalable and efficient alternative, where word embeddings are built incrementally by adding sparse random vectors each time a word co-occurs with a context lexeme. This approach is particularly appealing for modern distributional semantics because it allows embeddings to be constructed dynamically as data is processed, avoiding the computational overhead associated with pre-building matrices.

A related set of models to RI, also within the field known as Vector Semantic Architectures (VSA), is Holographic Reduced Representations (HRR), introduced by Plate in the 1990s (Plate 1994, 2003). Kanerva's Sparse Distributed Memory utilizes high-dimensional random binary vectors as addresses for storing data. The key idea is that similar inputs activate overlapping sets of memory locations, enabling content-addressable memory and pattern recognition. HRR, on the other hand, employs circular convolution and correlation operations to bind or associate vectors representing concepts. This creates a holographic-like representation where the whole (the composite representation) contains information about its parts (the constituent vectors), and the parts can be retrieved from the whole.

BEAGLE (Bound Encoding of the Aggregate Language Environment) extends the RI model by encoding word order information, also using circular convolution to bind context vectors in sequence. This model together with the HRR approach demonstrate the flexibility of random encoding models to handle both syntagmatic and paradigmatic relations effectively.

The final section on Self-Organizing Maps (SOMs) brings an interesting alternative use of random encodings to the table. SOMs apply competitive learning in a neural network architecture, reducing dimensionality while preserving the topology of the input space. Although SOMs are not new to the field of distributional semantics, the authors effectively link them to the family of random encoding models, emphasizing their ability to cluster lexemes based on contextual similarity. This draws attention to the broader potential of random encoding techniques in cognitive modeling and semantic learning.

By focusing on scalability, incrementality, and implicit dimensionality reduction, the authors highlight how these models address the limitations of earlier count-based models. The use of sparse, randomly generated vectors not only provides practical solutions to the challenges of processing large-scale corpora but also offers conceptual parallels to the way humans might incrementally learn semantics from context. Later in the text, they briefly explore the impact of random encoding models on the quality of semantic representations relative to newer neural-based models, such as those covered in the next chapter and Part Three.

In **Chapter Six**, after a brief introduction to the architecture of artificial neural network (ANNs), LS introduce the class of predict models known as neural language models (NLMs). They then proceed to explain the ability of such models to learn word embeddings as a by-product of predicting words. For example, given a sequence of words, the model predicts the most likely next word in the sequence. Predict models often produce higher-quality embeddings as they benefit from the predictive task's ability to refine word representations. They discuss the concept of a "static

embedding”, describing the two most influential architectures in the field: (1) Continuous Bag-of-Words (CBOW), which predicts a target word based on its surrounding context and aims to predict the middle word in a sequence by averaging the embeddings of context words; and (2) Skip-Gram, which predicts context words given a target word. It is particularly effective for learning from larger corpora, as it optimizes for predicting the surrounding words based on the target. Technically, both models learn embeddings (dense word vectors) via stochastic gradient descent (SGD) and optimize a loss function, typically negative log-likelihood, by minimizing the difference between predicted and actual context words.

One of the most innovative aspects of these approaches is the way in which the model learns the “labels” over the dataset without any annotation being necessary. This is known as self-supervised learning (although the training is technically supervised, in that input-output pairs are provided). The self-supervision comes from the fact that predicting the word in the middle (CBOW) or predicting the surrounding words (Skip-Gram) serves as a structured learning objective. The model does not need manually annotated labels because it can automatically derive this learning task from any raw text. Every word sequence becomes a training instance, with context words serving as inputs and target words serving as outputs.

Predict models have become standard as the input to contextualized learning approaches, discussed below. They are particularly advantageous because they can efficiently handle large datasets and do not need to maintain a complete co-occurrence matrix, allowing for scalable learning.

3.3. *The practice of DSMs*

Part Three begins with **Chapter Seven** discussing the key methodologies used to evaluate Distributional Semantic Models (DSMs), focusing on the established distinction between intrinsic and extrinsic evaluation strategies. Intrinsic evaluation focuses on how well DSMs approximate semantic similarity and relatedness as measured on specific tasks. It tests how distributional vectors capture various types of similarity using curated datasets. Tasks used for intrinsic evaluation have included: synonymy – evaluate the ability of DSMs to identify synonyms by measuring distributional similarity between word vectors. An example is the TOEFL synonymy test; similarity and relatedness – rate how similar or related words are; categorization – cluster words into predefined categories based on their vectors; analogy – evaluate an analogy using vector math, e.g., “man is to woman, as king is to ___”; relations – judging semantic relations, e.g., hypernymy, antonymy.

Extrinsic Evaluation involves testing DSMs on downstream NLP tasks, where their word vectors are used as features in machine learning models. The performance of the overall system (e.g., accuracy, F1-score) provides an indirect evaluation of the DSM’s quality. Such tasks include: Part-of-Speech Tagging (POS), Named Entity Recognition (NER), and Semantic Role Labeling (SRL). Here, distributional vectors are used as features in sequence labeling tasks; as well as document classification tasks. DSMs can also be evaluated on tasks requiring semantic understanding beyond single words, such as natural language inference and machine translation.

Interestingly, LS discuss Representation Similarity Analysis (RSA), a method developed in cognitive neuroscience, to compare the geometrical structure of the semantic spaces produced by different DSMs. RSA measures the similarity

between different DSMs by calculating the second-order similarity between their pairwise word similarities.

The challenges facing evaluation of DSMs involve data sparsity, task dependency, scalability, as well as instability of vectors and the influence of hyperparameters when training models.

The large-scale experiments demonstrate that while neural predict models often outperform traditional count models, the choice of context type and task-specific factors play a crucial role in model performance. The chapter also raises important questions about the instability of DSMs for low-frequency words and the limitations of current evaluation methods.

In **Chapter Eight**, LS turn finally to the application of distributional semantic models (DSMs) for representing aspects of linguistic theory (i.e., lexical semantics) that have been conventionally encoded by symbolic models. Symbolic models of lexical semantics represent meaning through abstract symbols structured into well-formed expressions by rules, as found in both formal semantics and cognitive semantics. Unlike DSMs, symbolic models are interpretable and explainable but struggle with capturing the graded and flexible nature of meaning in real-world language use, such as context sensitivity and vague meanings. The problem facing the distributional semanticist, on the other hand, is how to represent, in both a coherent and compositional manner, the information so compactly encoded in logical expressions or feature structures in symbolic models, while leveraging the flexibility introduced by continuous representations.

To illustrate the problem, consider an unstructured model of word meaning and how it might translate into a vector representation. Structuralist and generative decompositional models break down lexical meanings into primitive components or semantic features, which can be binary, such as those proposed in early Generative models of meaning (Chomsky 1965; Katz 1972; Katz and Fodor 1963):

car = [−ANIMATE, +ARTIFACT, +VEHICLE];
enter = [+DYNAMIC, +TELIC, −DURATIVE].

More complex structures can use conceptual categories and functions like Jackendoff's Conceptual Semantics or Generative Lexicon Theory, where the meaning of words like “book” is analyzed not only by its typing, but through Event and Qualia Structures, each structured as attribute-value pairs (Jackendoff 1990; Pustejovsky 1995). As the representational machinery for meaning becomes more complex, the challenge for distributional approaches to handle such expressions also becomes more daunting (cf. Palmer, Bonial, and Hwang 2017).

Similar remarks hold for relational models of meaning. These represent words by their relations to other lexemes rather than decomposing words into internal structures, as represented by named features or primitives. This can include both lexical networks like WordNet (Miller and Fellbaum 1991), where lexical items are nodes connected by relationships like synonymy, hypernymy, and meronymy; or semantic networks more generally, where any conceivable relation between nodes can be modeled in a directed graph as a labeled edge (Collins and Quillian 1969).

LS identify the core problems facing DSMs from the perspective of the lexicon: (i) word senses; (ii) paradigmatic semantic relations; (iii) cross-lingual representations;

(iv) connotative meaning; (v) semantic change; (vi) extralinguistic grounding; and (vii) cognitive modeling.

Early proposals for overcoming the brittleness of categorical representations often involve word meanings modeled as semantic feature vectors derived from an established inventory of dimensions (Smith and Medin 1981), or from features associated with the linguistic properties of a word in use (McRae et al. 2005). Since these are encoded as vectors, the tools available for vector similarity can be applied to measure the degree of meaning overlap or difference. More recent attempts simply create fully distributional vectors, where information is encoded into distributed patterns of dimensions.

In the context of representing word sense, LS address the core issue in lexical semantics of *polysemy*, the phenomenon where a single word can have multiple meanings depending on context (e.g., “bank” as a financial institution vs. a riverbank). Traditional models of the lexicon often treat word senses as discrete, dictionary-like entries. DSMs, however, provide a more dynamic and context-sensitive approach by clustering word usages based on distributional similarity (cf. Boleda 2020). DSMs can generate “word sense clusters” by identifying usage patterns in corpora. Each context of a polysemous word (e.g., *bank*) can be assigned a separate vector representation, and clustering methods (e.g., K-means, Gaussian Mixture Models) can be used to group similar contexts together (Erk 2012; Navigli 2009; Reisinger and Mooney 2010). For sense disambiguation tasks, DSMs use contextual embeddings to assign the correct sense of a word based on surrounding words. Transformer models like BERT work well in such cases because they encode words in their sentence-level contexts, dynamically adjusting word representations.

A major technical challenge in DSMs is the *meaning conflation deficiency* – the inability to cleanly separate different senses of polysemous words. While contextual embeddings like BERT mitigate this issue by assigning different vectors to the same word in different contexts, DSMs still struggle with generating discrete, interpretable senses without additional supervision. Technically, the key point is the transition from static embeddings (e.g., word2vec, where a word has a single embedding) to contextual embeddings (BERT, where each instance of a word has a context-dependent embedding), which better model polysemy.

LS then move to the ability of distributional models to distinguish between paradigmatic semantic relations, such as synonymy, antonymy, hyponymy, and meronymy. A key problem in DSMs is their tendency to capture general relatedness, without specifying the nature of the relation. Consider first synonymy and antonymy. In DSMs, similar words are clustered together based on context, but DSMs often fail to capture the *paradox of antonymy*, where antonyms (e.g., *hot* and *cold*) are semantically similar but denote opposite meanings. These words share a high degree of contextual overlap but represent opposing concepts.

Identifying taxonomic relationships (e.g., *dog* is a hyponym of *animal*) can also be difficult for DSMs because these are often hierarchical and require an understanding of generalization. The use of additional resources like WordNet can improve model performance in detecting these relations, but it requires supervised learning approaches. Supervised methods for relation classification are discussed, including techniques like relation extraction via dependency parsing and classification algorithms using hand-annotated datasets (e.g., WordNet for hypernymy). Techniques like *retrofitting* fine-tune DSMs to better capture specific relations by adjusting

vectors based on lexical constraints. Technically, DSMs rely on vector space models where relations are inferred through vector operations (e.g., offset methods), but these approaches often blur distinctions between related, but different, concepts. For example, DSMs might show *dog* and *animal* as related because they appear in similar contexts but not recognize that the relationship is hierarchical (hypernymy) rather than associative.

LS next turn to cross-lingual DSMs, which aim to align semantic spaces across languages, allowing for multilingual tasks like bilingual lexicon induction. Two main approaches are discussed: *mapping* models, where monolingual spaces are mapped into a common space (e.g., through bilingual dictionaries), and *joint models*, which learn monolingual and cross-lingual embeddings simultaneously. Both approaches rely on resources like parallel corpora and comparable corpora (e.g., Wikipedia) to align the vectors. A significant issue with current cross-lingual DSMs is their inability to handle meaning conflation, particularly for polysemous words, where different senses in one language may map to different lexemes in another (e.g., *bat* as animal vs. tool). Despite these challenges, cross-lingual embeddings represent a crucial step toward multilingual understanding in NLP.

LS address whether DSMs can represent connotative meanings, covering the emotional and evaluative aspects of words. Lexical items like *gang* and *group* differ not in denotation but in their emotional associations. The authors introduce the *Affective Distributional Hypothesis*, which suggests that a word's affective value can be predicted from the contexts in which it appears. Models like SentiWordNet and WordNet Affect are designed to capture these subtleties, though their coverage is limited. Importantly, LS also address cultural stereotypes and biases in DSMs, noting that the co-occurrence patterns that DSMs are built on can inadvertently encode and perpetuate stereotypes.

LS discuss one of the most challenging topics in the field, exploring how distributional models can be extended to include grounding through non-linguistic modalities, such as visual data or sensorimotor experiences. Grounded representations integrate linguistic and perceptual information, with the goal of providing a more complete model of meaning. Traditional distributional models, which learn word meanings based on textual co-occurrences, fail to bridge the gap between linguistic symbols and the physical world. Grounding linguistic representations in sensory experiences is essential for creating representations that align with human cognition. LS discuss *multimodal distributional semantics*, which combines linguistic data with other forms of data (e.g., visual, auditory, motor). This includes techniques that merge word embeddings with embeddings learned from other modalities like images or videos. One popular technique is a joint multimodal embedding, where the goal is to learn a unified vector space where both words and perceptual input are represented in the same space. This allows for a direct comparison of word meanings across modalities (e.g., linguistic descriptions of objects and images of those objects). Models are trained on paired data, such as image captions, where each word is associated with visual features. A common architecture for this is a shared projection space, where both text and visual embeddings are projected into a common latent space.

The technical implementation of grounded models typically involves neural network architectures that integrate different modalities. A key task for grounded distributional models is aligning textual and visual data by matching word vectors with

image embeddings. The model uses a cross-modal objective to maximize similarity between text and image embeddings. Deep learning models, like convolutional networks for visuals and recurrent networks for text, map each modality to a shared latent space, with a fusion layer combining these vectors.

The final section explores the cognitive plausibility of DSMs in modeling the mental lexicon and semantic memory. DSMs are compared to theories of semantic memory, noting their ability to predict human behaviors in tasks like priming and word association. The authors also discuss the alignment of DSMs with cognitive theories like the Strong Distributional Hypothesis, which suggests that meaning is determined by linguistic use. DSMs show promise in predicting human judgments, but the chapter closes by highlighting the need for further research to refine their cognitive models.

Chapter Nine broadens the application of distributional semantics from lexical items to complex linguistic expressions, such as phrases and sentences. The authors emphasize the importance of moving beyond the lexicon to address the compositionality inherent in natural language, arguing that for distributional semantics to be an expressive model of meaning, it must adequately represent and construct the meaning of these complex expressions.

Compositionality refers to the principle that the meaning of a complex linguistic expression (like a phrase or sentence) is determined by the meanings of its parts and the syntactic rules used to combine them. This principle, rooted in Fregean semantics, poses a significant challenge for distributional semantics, which traditionally excels in modeling word-level meaning but struggles to extend that to multi-word expressions. The chapter frames the challenge of extending DSMs to account for compositionality by referencing the Fregean notion: the meaning of a whole is derived from the meaning of its parts. Distributional vectors, while effective at capturing word-level meaning based on co-occurrence patterns, must be composed to capture the meaning of phrases and sentences. Hence the core problem can be stated as follows: how can we compose individual word vectors to represent more complex linguistic units like phrases, sentences, or even paragraphs, while preserving the syntactic and semantic dependencies between words? The challenge for distributional semantics is how to achieve compositionality using vector-based representations, where meaning is encoded in continuous spaces rather than logical forms.

Compositionality in distributional semantics is more than just summing or combining vectors. It also involves accounting for how word meanings shift and interact in context. For example, consider the noun *book* in isolation versus in the phrase *finish a book*, where “book” takes on a different aspect (as something to be read). A crucial point raised is that in symbolic models, word disambiguation typically precedes composition (word senses are fixed), while in distributional models, composition can influence meaning dynamically, requiring a mechanism for shifting and adjusting meanings during composition.

To this end, LS review the various methods for composing word vectors into phrase and sentence vectors. The simplest method involves summing or averaging the word vectors, but it has limitations, such as not encoding word order and treating all words equally, failing to capture syntactic relations. Multiplicative models are another method introduced, where vectors are combined elementwise. This approach introduces interactions between words, as the product emphasizes common elements, effectively performing implicit disambiguation by focusing on shared features.

However, the method still suffers from insensitivity to word order and is prone to information loss, as multiplying sparse vectors can result in too many zero values, reducing the expressiveness of the final sentence vector.

They also discuss tensor product models (Smolensky 1990), which capture word interactions more richly by producing a higher-order tensor rather than a simple vector. These models are more expressive and can account for word order and argument structure (e.g., *dogs chase cats* vs. *cats chase dogs* would produce different tensors), but the cost is an exponential increase in dimensionality, making them computationally expensive and hard to scale.

After reviewing the major vector-mixture models (where the types of what you are combining results in the same type of thing, i.e., vectors), the central focus of this chapter involves a detailed examination of new approaches to compositionality developed over the past decade or so, which LS group together as *Distributional Functional Models* (DFMs) (Baroni 2013; Baroni, Bernardi, and Zamparelli 2014; Clark, Coecke, and Sadrzadeh 2008). Parallel to the domain of types in simple type theory (STT) as employed in formal semantic models of language meaning, such as Montague (Sutton 2024), DFMs aim to incorporate the compositional structure of language into distributional models by representing: (a) sentences as vectors; (b) arguments (such as nouns or entities) as vectors; and (c) predicates (such as verbs, adjectives) as matrices or higher-order tensors. The intuition here is that sentences and nouns correspond to the “simple types” from STT, that is individuals and truth-values. While this takes some liberties with the underlying model theory of the type system – since nouns are actually treated like predicates in STT – (Baroni, Bernardi, and Zamparelli 2014), it allows for a formal distinction between simple objects (vectors) and more complex objects, called *distributional functions* (similar to functional types in STT).

Distributional functions have compositional properties that are similar to operations in STT. For example, predicates (such as verbs or adjectives) are represented by higher-order tensors or matrices, which operate on vectors (arguments). In DFMs, predicates are represented as linear transformations (e.g., matrices or tensors) that operate on noun vectors. For example, an adjective like “fast” is a matrix that modifies the noun vector “dog” to produce a vector representing “fast dog”. Higher-order predicates, such as transitive verbs, are represented by tensors that capture interactions between multiple arguments. A transitive verb like *chase* is modeled with a third-order tensor that composes with both subject and object arguments.

The DFM makes use of matrix-vector multiplication to capture the combinatorial nature of meaning composition, analogous to function application in formal semantics. The subject vector and the verb’s tensor interact to generate the verb phrase, which is then combined with the object vector. This parallels the way that symbolic logic builds interpretations through function-argument structures. STT’s functional types are typically context-independent once assigned, whereas distributional types are inherently context-sensitive, where the meaning of a word (and its type) can shift depending on the other words in its local context (e.g., the vector for “dog” might change in different syntactic and semantic contexts). The question addressed here has been how individual word embeddings are composed using formal operations, where sentence-level representations are built compositionally, and the aim has been to model the mechanisms of algebraic composition of the structure of phrases or sentences.

So far, we have focused on word embeddings and how they can be combined into larger representations. Our discussion above shows how sentences can be converted to vectors through matrix algebraic operations with distributional function application and composition. Distributional functions use methods such as: tensor product representations; matrix-vector multiplication; and linear combinations of vectors (e.g., additive or multiplicative models). These methods handle the compositional structure of a sentence but do not, by themselves, generate contextualized embeddings for entire sentences. We turn to this discussion next.

LS review some of the most recent developments in the field for sentence-level representations known as *sentence embeddings*. Sentence embeddings are vector representations of entire sentences, typically generated by neural network models (e.g., RNNs, Transformers) that learn to produce a single fixed-size vector to represent the entire sentence. For example, in models like BERT, the sentence “John runs” is encoded into a single vector (e.g., the vector for the [CLS] token), which represents the meaning of the entire sentence in a contextualized way. One advantage is that sentence embeddings provide single vector representations for entire sentences, capturing both word-level meanings and contextual dependencies between words. Unlike distributional functions, which rely on algebraic combinations of pre-trained word vectors, sentence embeddings are often learned representations created by processing entire sentences in sequence (or through self-attention mechanisms in Transformers).

LS discuss the primary models generating such embeddings. Recurrent Neural Networks (RNNs) are models that read sentences word by word and maintain a hidden state that evolves over time. The final hidden state is often treated as the sentence embedding. Transformer Models (e.g., BERT, GPT), on the other hand, use “self-attention” to compute dependencies between all pairs of words in a sentence. Sentence embeddings generated by models like BERT are context-sensitive. Each word’s vector in the sentence is dynamically updated based on the words around it, so the final sentence embedding captures not only the meanings of individual words but also how these words interact within this actual sentence. These aspects of sentence meaning thus learned during this training are held in a special token inserted for this purpose (e.g., the [CLS] token in BERT).

Transformers represent the current state-of-the-art approach for sentence embeddings, especially in models like BERT and GPT. The key innovation of transformers is the *self-attention* mechanism, which allows the model to compute dependencies between all pairs of words in a sentence, regardless of their distance from each other. This mechanism allows the model to focus on important words and relationships regardless of their position in the sentence, capturing both local and global dependencies. GPT generates text using a left-to-right (causal) approach and lacks a predefined token for sentence-level tasks. The last token’s hidden state can be used to capture the entire sentence’s information, though it’s not designed specifically for summarizing the sentence. In GPT, the closest analogue to BERT’s [CLS] token would be the hidden state of the last token in the sequence.

Since the introduction of LLMs, the importance of sentence embeddings has become obvious, for numerous NLP applications, from text classification and machine translation to chat and dialogue systems. As neural architectures continue to evolve and model sizes increase, the quality and applicability of sentence embeddings will likely improve dramatically.

Chapter 10 reflects on the achievements of distributional semantics, particularly in modeling lexical meanings and their application in NLP. However, significant challenges remain in areas such as compositionality, contextualized meaning, evaluation, and cognitive modeling. The authors advocate for future research to focus on developing hybrid models that combine the strengths of symbolic semantics with distributional methods, improving scalability, and addressing ethical and interpretability concerns.

4. Concluding remarks

Given that this book is largely concerned with how to interpret language data, I would like to conclude with some remarks on the nature of data in linguistics. Seventy years ago, linguistics had a data problem. There was not enough observable language data to fully operationalize the “discovery procedures” needed by structuralist methodologies and distributional analysis to induce or generalize the syntactic regularities inherent in the language. American structural linguistics, heavily influenced by Bloomfieldian principles, emphasized the collection and analysis of observable data – namely, corpus data derived from spoken and written language samples. In this context, Harris’s work on distributional linguistics occupies a pivotal position in the history of American linguistics, particularly in the transition from post-Bloomfieldian structuralism to generative grammar. At a time when linguistic theory focused predominantly on the systematic description of observable data, Harris applied structuralist principles with formal rigor. His observations on transformational families of constructions in language admitted degrees of abstraction, not only in the description but over the data being examined as well. At the time, his contemporaries may have viewed his methods as a refined, logical extension of what many were already practicing: deriving linguistic categories and functions from patterns in empirical data. In hindsight, however, Harris’s formalization of distributional methods is seen as groundbreaking. While firmly within the empirical tradition, Harris introduced mathematical techniques that anticipated generative grammar’s abstraction and theoretical models. His distributional methods helped bridge the gap between description and explanation, laying a foundational steppingstone toward uncovering the deeper structures of language – a move later formalized by Chomsky. Thus, even though Harris’s work may have appeared as a culmination of structuralist methods, it now stands as a key precursor to generative grammar’s explanatory shift (Newmeyer 2022).

But the data scarcity problem remained. Even with Harris’s formal developments in distributional analysis, the focus was on inductive methods: deriving linguistic rules from patterns found in empirical data. Linguists were cautious about incorporating native speaker intuitions, as these were considered subjective and potentially unreliable. Chomsky challenged this approach by arguing that native speakers possess an inherent knowledge of language, the grammatical rules of which can be accessed through their intuitions about sentence acceptability (Chomsky 1957). This marked a shift towards rationalist methodologies, where introspective judgments became valid and essential data for linguistic analysis. By legitimizing the use of intuitionistic data, Chomsky liberated linguistics in an era of data scarcity and enabled linguists to examine grammatical structures that may not appear in existing corpora but are nonetheless part of the language’s potential; to construct their own data (words,

phrases, sentences) to test specific theoretical predictions, allowing for a more rigorous evaluation of linguistic theories; and significantly, to investigate underlying representational structures that are not immediately evident from surface-level data (Tomalin 2006).

This approach was at the time groundbreaking, because it expanded the data set available to linguists beyond the limitations of corpora, which are inherently finite and may not capture the full range of a language's grammatical possibilities. Hence, the incorporation of intuitionistic data into linguistic analysis was a major revolutionary aspect of Chomsky's work during 1955–1962. It fundamentally changed the methodology of linguistic research by expanding the types of data considered valid and by emphasizing the mental reality of grammatical knowledge. In hindsight, however, it can be argued that Chomsky's continued insistence on treating corpus-collected linguistic generalizations as artifacts of behavior rather than reflections on mental competencies has failed to appreciate the larger intellectual shift in computationally informed linguistics.

Which brings us 70 years later to 2024. The availability of language data through mining and scraping webpage content has itself revolutionized the way linguistic data is used in linguistics. Contemporary linguistic data collections provide the source for remarkable constructional discovery and grammatical induction, all of which are being realized in current large language models. But the language sciences have another data problem: a glut of linguistic data and little appreciation of the cognitive states of the agents who have generated all this data. If we take a Popperian view of scientific discovery, theory construction and data analysis should not be strange bedfellows but work together (Popper 2005). Real data does more than merely provide support to the theories we have constructed; data is the missing ingredient to a theoretical approach to language as a science to begin with. This entails a cycle of examining the data with a critical but non-biased take on the phenomena. Theoretical fragments emerge but are incomplete and leaky. These are used as an examined lens over the data again, resulting in new observations, new theoretical constructions, and lo, new data. The data cycle begets new data through increased observational adequacy from new theoretical understanding.

And this is why Lenci and Sahlgren's book is important, as it presents an impressive treatment of the historical and theoretical foundations of distributional semantics. The compilation, synthesis, and communication of a complex and heterogeneous body of research can have significant impact on a discipline, particularly when the field is undergoing profound changes. By providing a clear and rigorous account of how meaning can be modeled through usage patterns, Lenci and Sahlgren contribute to ongoing debates in semantics and cognitive science dating back 70 years. Their book supports the view that meaning in language can be understood through its distributional properties, which can in turn help linguists better understand the formal properties inherent in language meaning.

Acknowledgements

I would like to thank Kyeongmin Rim, Bingyang Ye, Jingxuan Te, and Kenneth Lai, for suggestions provided on an earlier draft of this article. I would also like to thank the editor for substantive comments and edits.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

James Pustejovsky  <http://orcid.org/0000-0003-2233-9761>

References

- Baroni, Marco. 2013. "Composition in Distributional Semantics." *Language and Linguistics Compass* 7 (10): 511–522. <https://doi.org/10.1111/lnc3.12050>.
- Baroni, Marco, Raffaella Bernardi, and Roberto Zamparelli. 2014. "Frege in Space: A Program for Compositional Distributional Semantics." *Linguistic Issues in Language Technology* 9:241–346.
- Bloomfield, Leonard. 1933. *Language*. New York: Holt, Rinehart & Winston.
- Boleda, Gemma. 2020. "Distributional Semantics and Linguistic Theory." *Annual Review of Linguistics* 6 (1): 213–234. <https://doi.org/10.1146/annurev-linguistics-011619-030303>.
- Chomsky, Noam. 1957. *Syntactic Structures*. Berlin: De Gruyter Mouton.
- Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.
- Church, Kenneth, and Patrick Hanks. 1989/1990. "Word Association Norms, Mutual Information, and Lexicography." *Computational Linguistics* 16 (1): 22–29.
- Clark, Stephen, Bob Coecke, and Mehrnoosh Sadrzadeh. 2008. "A Compositional Distributional Model of Meaning." In *Proceedings of the Second Quantum Interaction Symposium (Qi-2008)*, edited by Peter D. Bruza, William Lawless, and Keith van Rijsbergen, 133–140. Saarbrücken: College Publications.
- Collins, Allan M., and M. Ross Quillian. 1969. "Retrieval Time from Semantic Memory." *Journal of Verbal Learning and Verbal Behavior* 8 (2): 240–247. [https://doi.org/10.1016/S0022-5371\(69\)80069-1](https://doi.org/10.1016/S0022-5371(69)80069-1).
- Constant, Mathieu, Gülşen Eryiğit, Johanna Monti, Lonneke Van Der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. "Multiword Expression Processing: A Survey." *Computational Linguistics* 43 (4): 837–892. https://doi.org/10.1162/COLI_a_00302.
- Erk, Katrin. 2012. "Vector Space Models of Word Meaning and Phrase Meaning: A Survey." *Language and Linguistics Compass* 6 (10): 635–653. <https://doi.org/10.1002/lnc3.362>.
- Firth, John R. 1951. "Modes of Meaning." In *Essays and Studies. Oxford: The English Association*. Reprinted in *Papers in Linguistics 1934–1951* (1957), edited by John R. Firth, 190–215. London: Oxford University Press.
- Firth, John R. (1957) 1968. "A Synopsis of Linguistic Theory 1930–55." In *Selected Papers of J.R. Firth (1952–59)*, edited by Frank R. Palmer, 168–205. London and Bloomington, IN: Longman and Indiana University Press.
- Hanks, Patrick. 2004. "The Syntagmatics of Metaphor and Idiom." *International Journal of Lexicography* 17 (3): 245–274. <https://doi.org/10.1093/ijl/17.3.245>.
- Hanks, Patrick. 2013. *Lexical Analysis: Norms and Exploitations*. Cambridge, MA: MIT Press.
- Hanks, Patrick, and James Pustejovsky. 2005. "A Pattern Dictionary for Natural Language Processing." *Revue Française de Linguistique Appliquée* 10 (2): 63–82. <https://doi.org/10.3917/rfla.102.82>.
- Harris, Zellig S. 1946. "From Morpheme to Utterance." *Language* 22 (3): 161–183. <https://doi.org/10.2307/410205>.
- Harris, Zellig S. 1951. *Methods in Structural Linguistics*. Chicago, IL: University of Chicago Press.
- Harris, Zellig S. 1954. "Distributional Structure." *WORD* 10 (2-3): 146–162. <https://doi.org/10.1080/00437956.1954.11659520>.
- Jackendoff, Ray. 1990. *Semantic Structures*. Cambridge, MA: MIT Press.
- Jezek, Elisabetta, and Patrick Hanks. 2010. "What Lexical Sets Tell us About Conceptual Categories." *Lexis: E-Journal in English Lexicology* 4: 7–22.

- Johnson, William, and Joram Lindenstrauss. 1984. "Extensions of Lipschitz Mappings into a Hilbert Space." *Conference in Modern Analysis and Probability* 26:189–206. <https://doi.org/10.1090/conm/026/737400>.
- Kanerva, Pentti. 1988. *Sparse Distributed Memory*. Cambridge, MA: MIT Press.
- Katz, Jerrold J. 1972. *Semantic Theory*. New York: Harper & Row.
- Katz, Jerrold J., and Jerry A. Fodor. 1963. "The Structure of a Semantic Theory." *Language* 39 (2): 170–210. <https://doi.org/10.2307/411200>.
- Landauer, Thomas K., and S. Thomas Dumais. 1997. "A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge." *Psychological Review* 104:211–240. <https://doi.org/10.1037/0033-295X.104.2.211>.
- Levy, Omer, and Yoav Goldberg. 2014. "Linguistic Regularities in Sparse and Explicit Word Representations." In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, Ann Arbor, MI, edited by Roser Morante and Scott Wen-tau Yih, 171–180. Stroudsburg, PA: Association for Computational Linguistics. <https://doi.org/10.3115/v1/W14-1618>.
- Lund, Kevin, and Curt Burgess. 1996. "Producing High-dimensional Semantic Spaces from Lexical Co-occurrence." *Behavior Research Methods, Instruments, & Computers* 28 (2): 203–208. <https://doi.org/10.3758/BF03204766>.
- Matthews, Peter H. 1993. *Grammatical Theory in the United States from Bloomfield to Chomsky*. Cambridge, MA: Cambridge University Press.
- McRae, Ken, George S. Cree, Mark S. Seidenberg, and Chris McNorgan. 2005. "Semantic Feature Production Norms for a Large Set of Living and Nonliving Things." *Behavior Research Methods* 37 (4): 547–559. <https://doi.org/10.3758/BF03192726>.
- Miletić, Filip, and Sabine Schulte im Walde. 2024. "Semantics of Multiword Expressions in Transformer-Based Models: A Survey." *Transactions of the Association for Computational Linguistics* 12:593–612. https://doi.org/10.1162/tacl_a_00657.
- Miller, George A., and Christiane Fellbaum. 1991. "Semantic Networks of English." *Cognition* 41 (1-3): 197–229. [https://doi.org/10.1016/0010-0277\(91\)90036-4](https://doi.org/10.1016/0010-0277(91)90036-4).
- Moon, Rosamund. 2007. "Words, Grammar, Text: Revisiting the Work of John Sinclair." *International Journal of Corpus Linguistics* 12 (2): 159–181. <https://doi.org/10.1075/ijcl.12.2.05moo>.
- Navigli, Roberto. 2009. "Word Sense Disambiguation: A Survey." *ACM Computing Surveys* 41 (2): 10:1–10:69.
- Newmeyer, Frederick J. 2022. *American Linguistics in Transition: From Post-Bloomfieldian Structuralism to Generative Grammar*. Oxford: Oxford University Press.
- Osgood, Charles E., George J. Suci, and Percy H. Tannenbaum. 1957. *The Measurement of Meaning*. Urbana, IL: University of Illinois Press.
- Palmer, Martha, Claire Bonial, and Jena D. Hwang. 2017. "VerbNet: Capturing English Verb Behavior, Meaning, and Usage." In *The Oxford Handbook of Cognitive Science*, edited by Susan E. F. Chipman, 315–336. Oxford: Oxford University Press.
- Plate, Tony A. 1994. "Distributed Representations and Nested Compositional Structure." PhD thesis, Department of Computer Science, University of Toronto.
- Plate, Tony A. 2003. *Holographic Reduced Representation: Distributed Representation for Cognitive Structures*. Vol. 150. Stanford: CSLI Publications.
- Popper, Karl. 2005. *The Logic of Scientific Discovery*. London/New York: Routledge.
- Pustejovsky, James. 1995. *The Generative Lexicon*. Cambridge, MA: MIT Press.
- Pustejovsky, James, Patrick Hanks, and Anna Rumshisky. 2004. "Automated Induction of Sense in Context." In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, edited by Sergei Nirenburg, 924–930. Geneva: COLING.
- Reisinger, Joseph, and Raymond 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*, edited by Ron Kaplan, Jill Burstein, Mary Harper, and Gerald Penn, 10–117. Los Angeles, CA: Association for Computational Linguistics.
- Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

- Smith, Edward E., and Douglas L. Medin. 1981. *Categories and Concepts*. Cambridge, MA: Harvard University Press.
- Smolensky, Paul. 1990. "Tensor Product Variable Binding and the Representation of Symbolic Structures in Connectionist Systems." *Artificial Intelligence* 46 (1-2): 159–216. [https://doi.org/10.1016/0004-3702\(90\)90007-M](https://doi.org/10.1016/0004-3702(90)90007-M).
- Sutton, Peter R. 2024. "Types and Type Theories in Natural Language Analysis." *Annual Review of Linguistics* 10 (1): 107–126. <https://doi.org/10.1146/annurev-linguistics-031422-113929>.
- Tomalin, Marcus. 2006. *Linguistics and the Formal Sciences: The Origins of Generative Grammar*. Vol. 110. Cambridge, MA: Cambridge University Press.
- Wells, Rulon S. 1947. "Immediate Constituents." *Language* 23 (2): 81–117. <https://doi.org/10.2307/410382>.
- Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. Edited by Gertrude E. M. Anscombe and Rush Rhees, translated by Gertrude E. M. Anscombe. Oxford: Blackwell.