TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 00 (2024) 1–31

ISSN: 1756-8765 online
DOI: 10.1111/tops.12771

This article is part of the topic "2023 Rumelhart Prize Issue Honoring Nick Chater," Morten H. Christiansen and Mike Oaksford (Topic Editors).

Distributional Semantics: Meaning Through Culture and Interaction

Pablo Contreras Kallens,[a] Morten H. Christiansen[b,c]

[a]*Department of Language Science and Technology, Saarland University*
[b]*Department of Psychology, Cornell University*
[c]*School of Communication and Culture, Aarhus University*

Received 29 March 2024; received in revised form 24 October 2024; accepted 28 October 2024

Abstract

Mastering how to convey meanings using language is perhaps the main challenge facing any language learner. However, satisfactory accounts of how this is achieved, and even of what it is for a linguistic item to have meaning, are hard to come by. Nick Chater was one of the pioneers involved in the early development of one of the most successful methodologies within the cognitive science of language for discovering meaning: distributional semantics. In this article, we review this approach and discuss its successes and shortcomings in capturing semantic phenomena. In particular, we discuss what we dub the "distributional paradox:" how can models that do not implement essential dimensions of human semantic processing, such as sensorimotor grounding, capture so many meaning-related phenomena? We conclude by providing a preliminary answer, arguing that distributional models capture the statistical scaffolding of human language acquisition that allows for communication, which, in line with Nick Chater's more recent ideas, has been shaped by the features of human cognition on the timescale of cultural evolution.

Keywords: Distributional models; Large language models; Semantics; Symbol grounding; Embodiment; Language learning; Cultural evolution

Correspondence should be sent to Morten H. Christiansen, Department of Psychology, Cornell University, 228 Uris Hall, Ithaca, NY 14853, USA. E-mail: christiansen@cornell.edu

History
Has turned into words
The world
Has turned into words
Everything
Has turned into words
Words, words, words in bulk
Silvio Rodríguez, "Palabras"

1. Introduction

How do words get their meaning? Philosophers have pondered this question for millennia, yet disagreement abounds.[1] Since Plato (*Cratylus*, 1999), some scholars have proposed that linguistic meaning is "derived" from the mind (e.g., Chomsky, 2015; Fodor, 2001; Langacker, 2008). When someone says, "the cat is on the mat," the meanings of "cat" and "mat" depend on the listener's internal concepts of CAT and MAT, respectively. Other philosophers have instead argued that the meaning of a word comes from its usage in the language (e.g., Wittgenstein, 1953). So, the meanings of "cat" and "mat" come from the specific situational context in which they are being used. The debate has also spilled over into cognitive science, where both perspectives are represented. Nick Chater, too, entered the fray early in his career, conducting some key pioneering work with Steven Finch and Martin Redington, within what has since then become known as "*distributional semantics."*

Distributional approaches to semantics are based on two core intuitions about meaning. The first one is Firth's (1957) insight that the meaning of a word constrains what other words co-occur with it. Thus, the *context* of a word can be used as a window[2] to its meaning. The second one is that word meanings can be construed as a semantic *space* representing the similarity between those words in terms of their meaning (Salton et al., 1975). Even if we cannot express what exactly the meaning of "dog" is, we know it is more similar to "cat" than it is to "building," and "building" itself is more similar to "house" than to "run," and so on.

Both intuitions inspired the methodology behind distributional models of semantics: the meaning of a word can be captured by tracking its patterns of occurrence (hence, its *contexts*), and similar words will have similar patterns of occurrence. Thus, a *meaning space* can be built based on how words co-occur with one another, such that words that occur in the same contexts are closer to each other. The space itself will then be a model of the meaning of the words used to build it (Lund & Burgess, 1996).

In this article, we first discuss the contributions of Chater, Finch, and Redington to the nascent cognitive science approaches to distributional semantics in the early 1990s. This work focused on using word co-occurrences to capture broad aspects of word use in the form of lexical categories. We then survey subsequent computational models of increasing sophistication—including recent large language models (LLMs)—that can account for more detailed dimensions of word meaning and associated human behavior. These models, however, are not without limitations, which we discuss, including ways in which they do

not capture certain semantic phenomena, and various tentative attempts being undertaken to improve them. Finally, we return to our initial question about where the meanings of words come from and ponder what distributional semantics is really a model of, what their relation to language use could be, and what can be expected of them as explanatory tools.

2. Discovering lexical categories

Lexical categories, such as nouns and verbs, can be construed as broad semantic categories. Indeed, in English, nouns typically map onto objects/things and verbs to actions. So, knowing the lexical category of words provides (probabilistic) cues to their general meaning as well as how they might be used in a sentence. For example, knowing that "cat" and "mat" are nouns, "sat" a verb, "on" a preposition, and "the" a determiner, allows a speaker to combine them into the sentence frame *Det Noun Verb prep Det Noun*, yielding "The cat sat on the mat."

Starting in 1991, Nick Chater and his colleagues published a series of computational studies demonstrating how a distributional learner might gain knowledge about the lexical category of a word from its pattern of occurrence with other words. An initial hybrid model by Finch and Chater (1991) employed statistical analyses to obtain bigram frequency patterns (for pairs of words) from a corpus consisting of 33 million words from USENET newsgroups to train a neural network via Hebbian learning. Subsequent clustering of word similarity measures derived from the network revealed groupings of words into lexical categories. As shown in Fig. 1, these categories are more fine-grained than typical lexical categories, for example, dividing proper nouns into countries and names and verbs into different inflectional categories. Finch and Chater (1992) extended this work with as a Hebbian network implementing unsupervised clustering that captured both standard lexical categories but also more semantically nuanced ones. In further USENET hierarchical cluster analyses, Finch and Chater (1994) showed that proto-phrases and even short sentences can be derived from word co-occurrence statistics. And Redington, Chater, Huang, Chang, Finch, and Chen (1995) demonstrated that the same approach works for French, German, and Mandarin Chinese.

Whereas the modeling by Finch, Chater, and colleagues was primarily concerned with showing that there was sufficient information in large swaths of language to discover lexical categories of differing granularity, Redington, Chater, and Finch (1993) were interested in whether children might be able to learn lexical categories from the speech they hear from parents and other caregivers. So, instead of using written text from USENET, Redington et al. used 2.1 million words of child-directed speech from the CHILDES database (MacWhinney & Snow, 1985). To further increase psychological validity, they also adopted a simpler way of capturing distributional regularities (inspired by the Hebbian network in Finch & Chater, 1992). They focused their analyses on the 1000 most frequent words in the corpus and recorded the co-occurrence of each of those words with 150 "context words." These context words comprised the top-150 most frequent words and their co-occurrence within two words before and after the target words was recorded (i.e., a five-word window was moved through the entire corpus—CONTEXT_{-2} CONTEXT_{-1} TARGET CONTEXT_{+1} CONTEXT_{+2}—to record context-target word co-occurrences). Each target word was then represented by a

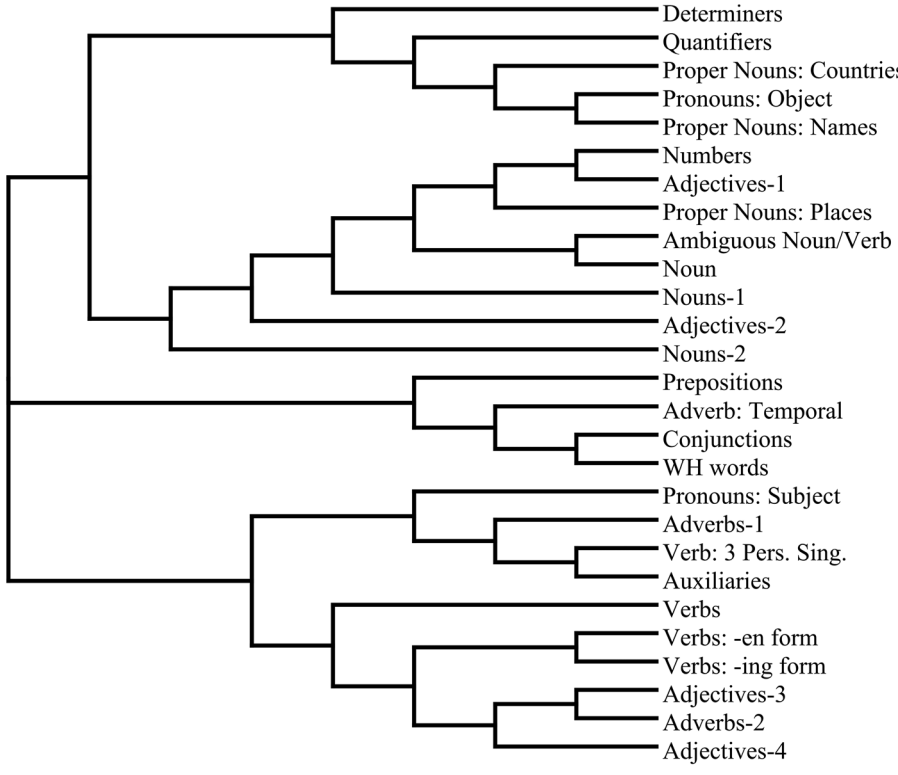

Fig. 1. Reproduction of a dendrogram from Finch and Chater (1991) showing the cluster structure of 1000 words, revealing a fine-grained lexical category structure (<5% of data was omitted or did not fit with the labels). This version of the dendrogram shows the same cluster structure as the original without reproducing the length of the branches. Constructed using Mesquite (Maddison & Maddison, 2023).

600-dimension vector corresponding to its co-occurrences with the context words. These vectors can then be compared to reveal similarities between words: Clustering of such word representations resulted in the discovery of lexical categories closely resembling those found by Finch and Chater (1991) but based on much less and more messy input. Moreover, Redington et al. showed that their model was able to classify new words as either nouns or verbs better than chance. Further distributional analyses in Redington, Chater, and Finch (1998) revealed that distributional information was more useful for classifying content words (nouns, verbs, etc.) than function words (determiners, prepositions, etc.). Redington and Chater (1998) additionally showed that within a lexical category cluster, there would sometimes be subclusters of words that were related semantically, as illustrated in Fig. 2, though in many other cases, the semantic relatedness was not high.

A possible limitation of Chater and colleagues' distributional analyses was that distributional information alone may not be sufficient for the discovery of lexical categories, let alone word meaning. To remedy this shortcoming regarding lexical categories, Monaghan, Chater, and Christiansen (2003) showed that phonological information can be a useful additional cue

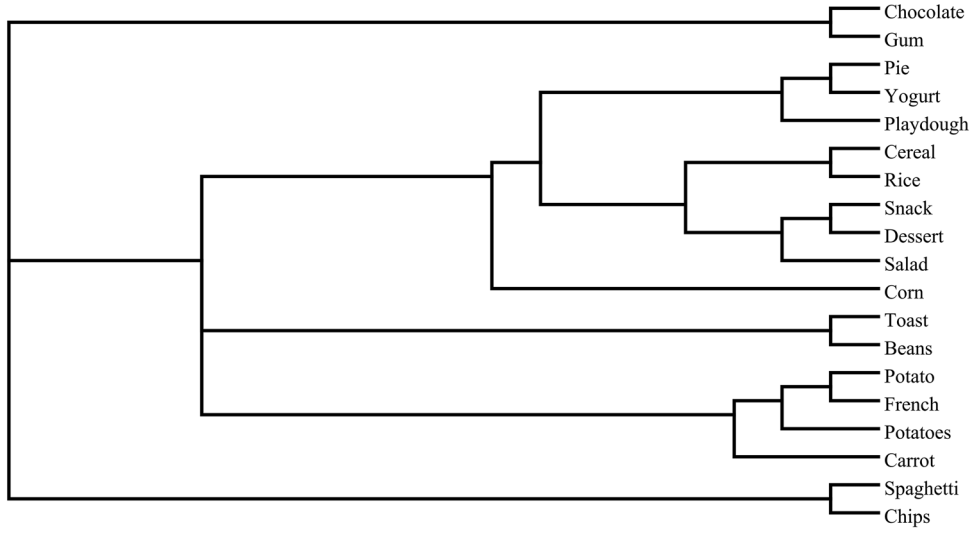

Fig. 2. A semantically related subcluster of nouns referring to food items, reproduced from Redington and Chater (1998). Note that although "playdough" is not an actual food, it is often consumed by small children (despite parents' best efforts) and thus appears in similar distributional contexts in child-directed speech. This version of the dendrogram shows the same cluster structure as the original without reproducing the length of the branches. Constructed using Mesquite (Maddison & Maddison, 2023).

to lexical categories, influencing human lexical processing. In subsequent work, quantifying the usefulness of both distributional and phonological cues to lexical categories, Monaghan, Chater, and Christiansen (2005) found a frequency by cue tradeoff whereby distributional cues were more useful for high-frequency words and phonological cues were more useful for low-frequency words. Monaghan, Christiansen, and Chater (2007) extended this line of work cross-linguistically to Dutch, French, and Japanese[3] (while also showing that the phonological-cue results were not driven by morphology).

The more ambitious project at the base of this, that is, accurately capturing the specific meaning of words through their distributional patterns, required methodological and architectural innovations. These innovations, in turn, show the potential behind the intuition of Nick Chater and his colleagues that an essential aspect of language acquisition can be attributed to tracking statistical patterns of use and abstracting from them.

3. Searching for meaning beyond lexical categories

Nick Chater's early work with his colleagues on how distributional information may be useful for language learning has helped pave the way for future work on distributional semantics. Note, however, that the focus on most of the subsequent distributional modeling was not on simulating the mechanism behind language learning in humans, but rather on effectively and efficiently inducing a generalizable representation of the meaning of words from given

textual data. In terms of their basic computational means of induction, these models can be roughly divided into "count" and "prediction" models (Baroni et al., 2014b; Lenci, 2018). We suggest that these distributional models can also be seen as milestones on the road to the current LLMs. Although the representation of meaning in LLMs is still not well understood, we nonetheless argue that they represent a further key step toward realizing the idea of meaning-through-use. Our discussion will focus on the conceptual proposals regarding what the context of a word is and how to induce meaning from it instead of the algorithmic details of their implementation—except for those cases where one is inextricably tied to the other.

3.1. Count-based models

The mechanism through which count-based models build their meaning space is by explicitly keeping track of the words that appear in the context of each word in the vocabulary. Count-based models can thus be viewed as generalizations on the distributional models used by Nick Chater and colleagues. After computing these *co-occurrence* statistics, some of these models include an additional step where they abstract a further, second-order space that captures the patterns present in the first-order, word co-occurrence space.

One of the earliest count-based models is the Hyperspace Analogue to Language (HAL; Lund & Burgess, 1996). This model takes an *occurrence window* approach to context: in a specific instance of language use, a word's context is the words that appear immediately before and after it. The occurrences are used to keep a count, for each word, of the times every other word has appeared as its context which are then tabulated into a co-occurrence matrix. The meaning of a word is represented as its row vector, that is, its co-occurrence with the rest of the words. This sparse matrix is truncated by considering only a subset of the words, such as those with the highest variance in context occurrence. The similarity between two words is measured by the Euclidean distance between their row vectors. The resulting truncated meaning space displays global similarity patterns such that, for example, words for countries (Lund & Burgess, 1996), alcoholic beverages (Burgess & Lund, 2000), and body parts (Lund & Burgess, 1997) form separate coherent clusters. Moreover, Burgess and Lund (2000) found that their similarity measure correlated with reaction times in a semantic priming study.

Latent Semantic Analysis (LSA; Landauer & Dumais, 1997) is perhaps the best-known distributional semantics model in cognitive science, at least until very recently (Jones et al., 2015). LSA started as a method for retrieving documents from queries to a database (Deerwester et al., 1990). However, Landauer and Dumais (1997) later repurposed the model as a method for extracting semantic knowledge from a corpus and representing the meaning of words. Instead of a local context of immediately surrounding words, LSA tracks the occurrence of words in separate documents across a corpus, creating a word $\times$ document matrix. Singular Value Decomposition is then applied to reduce the dimensionality of the matrix, keeping only the dimensions that capture the largest amounts of variance, ranging from just a few (e.g., 10, Contreras Kallens & Dale, 2018) to several hundred (Landauer & Dumais, 1997; see Evangelopoulos et al., 2012, for a range of useful dimensions). Conceptually, this step transforms a co-occurrence matrix into one that represents the *patterns* of co-occurrence. Landauer and Dumais (1997) found that LSA mapped onto human semantic

priming and ratings of meaning similarity; even scoring on par with non-native applicants to English-speaking universities (64.4%) when tested on the synonymy questions from TOEFL.

A third count-based approach, Topic Modeling, also focuses on word occurrences in documents—similar to LSA—but uses a Bayesian method for reducing dimensionality known as *Latent Dirichlet Allocation* (LDA; Blei et al., 2003). LDA is a generative probabilistic model that conceives a document as the result of a mixture of distributions over words, each document of a corpus being generated by the same set of topics in different proportions. These latent distributions over words are the *topics* of the corpus (Griffiths et al., 2007; Steyvers & Griffiths, 2007). Topics are operationalized as distributions over the probability of occurrence of each word in the corpus such that the word "dog" has a high probability of appearing in the topic "pets" but low on the topic "vehicles." Each document has a "gist" (Griffiths et al., 2007), which represents the particular mixture of topics that characterizes it. Griffiths et al. (2007) found that fitting the model with 1700 topics to the TASA corpus produced topics discernably related to theatre ("stage," "playwright"), scientific experimentation ("hypothesis," "evidence"), courts ("witness," "attorney"), school ("homework," "teacher"), and Marxist literature ("socialism," "revolution"). They furthermore showed that their model accounted for the same human data as LSA.

3.2. Prediction-based models

Whereas the previous models track the context in which a word occurs by counting their occurrences, prediction-based models achieve this implicitly by optimizing the representation of each word for a prediction task in a local linguistic context. Of these, word2vec (Mikolov et al., 2013a) is the best known one, with the paper originally proposing it accruing over 44,000 citations since its publication. The Hebbian network used by Finch and Chater (1992) to discover lexical categories can be seen as an early and more limited predecessor to the word2vec approach. A shallow neural network is trained to either predict a word based on the sequential presentation of its context (Continuous Bag of Words, CBOW) or the context based on the word (Skip-gram) (Mikolov et al., 2013a). The word2vec network repeatedly performs this task for the whole training set using gradient descent to optimize the weights of the connections between the input words and a linear hidden layer that then outputs to a prediction of the context (or vice-versa for the Skip-gram model). The connection weights between each word and the hidden layer after training are the vectors representing the meaning of each word.

Both methods for training word2vec are highly successful in modeling the semantic space of words. For example, Mikolov et al. (2013a) report that, when trained on a Google News corpus, it performs well on sentence completion tasks and can accurately reproduce close associates of common words. However, its most relevant feat is its ability to model word manipulation through simple vector combinations. For example, the result of the vector operation "King – Man + Woman" is a vector similar to the one for "Queen" (Mikolov et al., 2013a, p. 2) and the sum of "Russia" and "River" is a vector similar to "Volga" (Mikolov et al., 2013b, p. 7).

An important advance in prediction-based models was the introduction of the N-gram Skip-gram algorithm (Bojanowski et al., 2017), which instead of one-hot vectors for words uses a sum of the n-grams that form the word. This modified version significantly improves on word2vec's performance in word analogy tasks, with the largest improvement stemming from morphological analogies (Bojanowski et al., 2017). Crucially, the N-gram skip-gram model can represent and accurately generalize its space to rare and novel words by averaging the vectors of its component n-grams. This produces interesting patterns such as "scarceness" being related to "rarity" through the match of the n-grams composing "scarce" with the n-grams composing "rar" and the ones from "ness" matching with "ity" (Bojanowski et al., 2017, p. 11).

3.3. Large language models

A recent development in the induction and representation of meaning through statistics is the extraction of representations from more or less general-purpose language models (Devlin et al., 2019; Peters et al., 2018a; see Elman, 1991, for an early connectionist exploration of this idea). Recent innovations in model architecture, such as the Transformer (Vaswani et al., 2017), have made this approach more viable. In contrast with word2vec, which uses a shallow network trained on a specific task, LLM-based models involve networks with many layers, trained on the general task of predicting the next element ("token") in a sentence. The meaning of the target item is extracted from the activation of the network when the item is used as the input. Here, the context of each word, in the sense used by Firth (1957), is captured implicitly during the task of learning the sequential dependencies in the training corpus. Importantly, because the activation of the network for each item is dependent on the sequence of preceding items, the representation for a word depends on the context in which it was presented. Thus, the representation of the word "ape" in the context "the ape is eating the banana" will be different than the context "the fighter went ape mode" (Liu et al., 2020). This inherent context-dependence presents a new challenge to how the meaning of single terms can be effectively captured from the other parts of the input provided to the network (Bommasani et al., 2020; Lenci et al., 2022; Vulić et al., 2020). An example of the LLM approach to capturing meaning can be seen from the use of BERT (Devlin et al., 2019), where items are embedded into a semantic space by obtaining some combination of the activation of the layers after the input, be it only one or a weighted sum of a number of them. Devlin et al. (2019) found that different combinations can yield token representations with different properties. Other models such as GPT and its consecutive iterations (Radford et al., 2018; see Neelakantan et al., 2022, for an added optimization step), or T5 (Raffel et al., 2020), and, in principle, any neural language model (Mars, 2022) can be used to obtain these representations.

The representations of single words, regardless of their context, obtained through LLMs have been found to outperform the ones obtained by count-based and prediction-based methods (e.g., Bommasani et al., 2020; Ethayarajh, 2019; Peters et al., 2018b). Moreover, the in-context representations are able to capture different senses of the same word (Du et al., 2019; Wiedemann et al., 2019) while retaining coherence in its overall meaning (Coenen

et al., 2019; Ethayarajh, 2019). Crucially, because these meanings are based on extracting patterns of occurrence and generalizing them in ways that are *useful for language modeling*, the weights of the model, and thus the semantic space in which meanings are represented, have implicit knowledge of dimensions that determine usage such as syntax (Linzen & Baroni, 2021), morphology (Edmiston, 2020), and even some aspects of pragmatics (Potamias et al., 2020).

4. How close does distributional semantics come to human meaning?

The early computational modeling work by Nick Chater and his colleagues was motivated by the promise of how distributional information could provide human learners with important cues to use words appropriately (in terms of their lexical category). Subsequent work on distributional semantics has successfully expanded the models' reach dramatically to deal with many fine-grained aspects of meaning. Here, we discuss how well current models capture meaning-related human behavior and cognition. Through their limitations, they also provide a window into the aspects of meaning that are difficult to learn purely through tracking statistics of occurrence in use.

We have already touched upon the perhaps biggest success of distributional semantics, and the reason they can be considered as candidates for models of meaning in the first place: their robust correlation with human judgments about the similarity of words (see, e.g., Baroni et al., 2014b; Griffiths et al., 2007; Landauer et al., 1998; Pennington et al., 2014; Wang et al., 2019). Further comparisons of distributional methods with priming data (Jones et al., 2006; Mandera et al., 2017; Pereira et al., 2016) have confirmed that distributional semantics models can accurately predict whether semantic priming will occur, the strength of the effect, and general qualitative patterns with different kinds of words. LMM-based models have also been found to mirror human similarity when comparing pairs of sentences (see Chandrasekaran & Mago, 2021, for a review) and word senses in different contexts (Haber & Poesio, 2021; Sathvik Nair & Meylan, 2020). In what follows, we discuss work that deals with more specific psycholinguistic data, beyond the local geometric relations present in these spaces.

4.1. Modeling the acquisition of meaning

Nick Chater and colleagues' work with child-directed speech (Monaghan et al., 2005, 2007; Redington et al., 1993, 1998) provided early demonstrations that adult speech to children is sufficient for discovering lexical categories. Riordan and Jones (2011) found that most modern distributional models achieved representations of word classes similar to those contained in the MacArthur-Bates Communication Development Inventories (Fenson et al., 1994) when trained on the CHILDES (MacWhinney, 2014) corpora, rivaling hand-coded feature representations. Frermann and Lapata (2016) reproduced this result by training a modified Topic Model that includes word categories. Importantly, their model, trained incrementally, showed a developmentally plausible learning trajectory. This successful modeling of child word category acquisition was also achieved by prediction-based models (Asr et al., 2016; Huebner &

Willits, 2018). Wang et al. (2023) trained a long short-term memory (LSTM) model on the dense language input to just one child and found that embeddings extracted from this network clustered in a similar way.

Another developmentally relevant feature of distributional models is their ability to capture the dynamics of meaning acquisition in children. Landauer et al. (2011) developed the notion of "word maturity," measuring the difference in the representations of the same word between two different corpora. In their study, they built an LSA space with a complete corpus of paragraphs and compared this "final" state of the word's representation with its vector in spaces built with increasingly large portions of the corpus. They found that test words show an intuitive developmental trajectory: whereas words like "dog" and "turkey" maturate very quickly, others like "electoral" and "productivity" require most of the corpus (Landauer et al., 2011). In a more direct test, Biemiller et al. (2014) found that the LSA word maturity of a term has a higher correlation to Age of Acquisition than its print frequency.

Similarly, for prediction-based meaning representations, Botarleanu et al. (2021) showed that an analog of word maturity using word2vec in a set of multilingual corpora was a significant predictor of Age of Acquisition in the respective languages. Finally, Fourtassi et al. (2019) represented lexical development as a network where the edges between terms are derived from the cosines between their vectors in the word2vec embeddings. They found that the emergence of the cluster structure of this network on a month-by-month basis can mirror the real developmental trajectories of children. Whether LLMs also mirror these trajectories is not as well understood, and there is conflicting evidence. For example, Portelance et al. (2023) found that the average surprisal of words across their contexts in child-directed speech as computed by an LSTM is a significant predictor of Age of Acquisition beyond unigram frequency. By contrast, Chang and Bergen (2022) found that the average surprisal for each word during training (on a different corpus), treated as a proxy for its acquisition trajectory, does not have the same predictors as Age of Acquisition in children.

A final source of probabilistic information relevant for uncovering the meaning of words can be found in the sound patterns of words themselves (for reviews, see Dingemanse et al., 2015; Haslett & Cai, 2024). Indeed, Nick Chater's work with Monaghan and Christiansen had shown that child-directed speech contained useful information about lexical categories (Monaghan et al., 2003, 2005, 2007). Recent work has shown that distributional models are not limited to information available in word co-occurrence statistics but can also be made sensitive to such within-word cues to meaning. For example, when Gatti et al. (2023) provided their model with subword information, it was able to capture aspects of phonological patterning that affect human responses to nonwords in a semantic priming task. Moreover, when a multimodal LLM—trained on text and images—was tasked with generating 3D renderings of objects whose shape was defined by a nonword, it created shapes that were spiky for words sounding similar to *kiki* and rounded shapes for wounds that sounded like *bouba* (Alper & Averbuch-Elor, 2024). This replicates the well-known sound symbolic effect observed cross-linguistically in humans (e.g., Ćwiek et al., 2022). Future work exploring the degree to which distributional models take advantage of cues that complement word-level co-occurrence statistics can help inform our theories of how different sources of statistical

information are integrated, a process that is particularly important during the early stages of the acquisition of meaning (Imai & Kita, 2014; Monaghan et al., 2007).

4.2. Modeling brain activity

The distributional organization of words into classes has also been used as a predictor of brain activity related to specific meanings. Mitchell et al. (2008) built a distributional model using a set of concrete nouns and hand-selected verbs relating to basic sensorimotor activities. This model encoded the associative strength between each noun and verb by observing their co-occurrence in a corpus. Then, a classifier was trained to predict functional magnetic resonance imaging (fMRI) data using a linear combination of the vectors for the nouns. The classifier achieved an accuracy of 75% in a cross-validation scheme, showing that the co-occurrence-based semantic space was able to represent the similarity in brain activation through the similarity of the words. Pereira et al. (2011) reproduced these results using a Wikipedia-trained Topic Model, and Huth et al. (2016) used a co-occurrence matrix with a set of basic words from Wikipedia to construct a semantic map of the brain, showing what regions are activated when processing particular meanings.

All these studies, however, used only concrete nouns as stimuli for their models. Apart from the technical limitations of fMRI technology (Bullinaria & Levy, 2013), this is related to a problematic aspect of distributional models: the similarities being captured are too vague to account for the complete grasp of meaning that humans seem to possess. Indeed, Pereira et al. (2016) argue that pre-LLM distributional models are ill-equipped to deal with relationships of meaning that go beyond a vague notion of "relatedness." Evidence of this comes from the poorer performance of all models on the SimLex word similarity norms (Hill et al., 2015), which distinguishes between related words ("car" and "street"), semantically similar words ("car" and "automobile"), and both ("car" and "truck"). The inclusion of this control—meaning that a considerable portion of the norms are semantically similar but not related—drastically reduces the performance of all pre-LLM distributional models (Pereira et al., 2016). This overall difficulty of grasping subtler relationships of meaning is further reflected in the sharp decline of the correlation between model and human predictions when using norms that only include verbs (Gerz et al., 2016).

This problem has been tackled by using more sophisticated methods to construct a semantic space for encoding and decoding brain activity. Pereira et al. (2018) propose that prediction-based models such as word2vec can be used to build an effective "universal decoder" of semantic processing activity in the brain. Subsequent comparisons of these methods with LLM-based spaces have shown that the latter are superior due to properties that are in turn informative of the dimensions of language processing missed by the former. For example, LLMs are better at capturing the granularity of the brain's semantic space (Sun et al., 2020) thanks to the inclusion of context (Goldstein et al., 2022; Jain & Huth, 2018) and the optimization of the space for in-context prediction (Caucheteux & King, 2022; Goldstein et al., 2022). In other words, these models benefit from being the product of *optimization for language use* (although see Antonello & Huth, 2023, for a version of this argument that does not depend on prediction as a shared computational principle).

5. Have distributional models broken containment?

To discuss one of the key problems of distributional semantic models, especially in their count- and prediction-based forms, consider the assumption that the semantic space is homogeneous, that is, that there is no class or category of word that has a qualitatively different representation from the other ones. However, the semantic space of humans seems to be organized around sharper distinctions than this would suggest, for example, the difference between concrete and abstract words. Although it is a robustly attested phenomenon that more concrete and imageable words are easier to process than more abstract words (Kousta et al., 2011; Paivio, 1991), this difference is not easily captured by co-occurrence because it relates to the content rather than the context of a word.

The lack of differentiation in the space for the format and content of the representations—such as for abstract and concrete nouns—brings us then to two related but distinct objections to distributional semantics: the embodiment problem and the symbol grounding problem.

5.1. Language embodiment

The problem of the homogeneity of the semantic space in distributional models of meaning is consistent with the Embodied Cognition criticisms of distributional semantics (Glenberg & Mehta, 2012; Glenberg & Robertson, 2000; Zwaan, 2014). According to these perspectives, human cognition is inherently situated in the bodily experience of each agent—and this includes language processing and representation (Meteyard et al., 2012). In terms of distributional models, this would entail that representing words with radically different meanings using the same amodal disembodied symbols (Barsalou, 2016) cannot capture the way humans organize their knowledge.

One of the main claims of the embodiment perspectives on linguistic semantics is that understanding the meaning of a word involves activating the bodily experiences—perceptual and proprioceptive—related to that word (Barsalou, 1999; Zwaan, 2003). Thus, papering over the details, the meaning of "cat" is rooted in simulating perceptual experiences related to interacting with cats, and the meaning of "run" activates (among others) the proprioceptive experiences of running (Andrews et al., 2014; Binder & Desai, 2011). Depending on whether the embodiment is taken weakly or strongly (see Meteyard et al., 2012), simulation is necessary and/or sufficient for meaning beyond the associated words, respectively. Thus, from an embodiment perspective, purely distributional models of meaning are either inherently incomplete or entirely misguided.

The involvement of embodied processes for meaning can explain the problem of homogeneity as illustrated by the difference between abstract and concrete words. For example, according to some theorists (e.g., Barsalou et al., 2008; Borghi et al., 2017), the main difference between concrete and abstract meanings is their degree of embodiment. Concrete meanings activate more modality-specific areas of the brain (Montefinese, 2019) which suggests that their representation involves more sensorimotor simulation. Moreover, it may also explain the difficulties that distributed models tend to have with verbs: activation of action words involves activation of the motor systems responsible for performing it (Aziz-Zadeh

et al., 2006; see Vigliocco et al., 2011, for a review). But this information is less publicly available than the perceptual experiences associated with nouns, which could result in less informative co-occurrence relations in the records of public language use on which distributional models are trained (i.e., corpora).

5.2. The symbol grounding problem

The criticisms from embodiment are tightly tied to the problem of symbol grounding (Harnad, 1990). Although most reviews and criticisms treat these problems as largely equivalent (e.g., Andrews et al., 2014; Barsalou, 2016; Jones et al., 2015; Lenci, 2018), their main objections differ. Whereas embodiment criticisms focus on the implausibility of the *content* and *format* of distributional semantics, the objection of grounding focuses on their *usefulness* as models of meaning.

The basic intuition behind the grounding objection is the following. Imagine your knowledge of the meaning of words is totally captured by word2vec embeddings. Someone tells you "look, a cat!", and you attempt to understand what the meaning of the term "cat" is. You search your knowledge and find that the words closest to "cat" are "cats," "dog," and "kitten." Then, you access the most similar one, "cats," and find that its most similar words are "felines," "cat," and "pets." In turn, the most similar words to "felines" are "cats" and "feline," which brings you back to where you started.

This is the situation that Harnad referred to as the "symbol merry-go-round" (1990, p. 340): even though there are intuitive similarity relationships, the symbols do not seem to touch reality at any point. In more philosophical terms, the grounding problem of distributional semantic models is that they can be described as senses that do not fix any reference (Frege, 1948/1892; Putnam, 1974). The knowledge that they have about the meaning of "cat" does not enable them to identify a cat. Indeed, the relationships they reveal are only intuitive to human interpreters, which, of course, reveals that the knowledge of meaning that the interpreters have must be qualitatively different from the one that the model possesses (Searle, 1980). Therefore, the model cannot be an accurate description of the interpreter's knowledge of the meaning of words.

The reason the grounding problem has been tied to the embodiment objection is that embodying meanings can be seen as an answer to the question of how they can be grounded. Indeed, if the meaning of words is stored in concepts that inherently have sensorimotor properties, reference comes relatively cheaply (Barsalou, 1999). The meaning of words is tied to perceptual and proprioceptive features, and thus reference can be fixed with their mediation. Indeed, this was the solution to the symbol grounding problem that Harnad (1990) envisaged, based on early connectionist modeling (for a discussion of the philosophical difficulties facing this approach, see Christiansen & Chater, 1992, 1993).

Although still in their infancy, some recent distributional models have attempted to incorporate sensorimotor grounding into their representation. For example, some of them have attempted to integrate perceptual feature lists with Topic Models by training the generative model on both sources of data in parallel (Andrews et al., 2009, 2014; Steyvers, 2010). These models can predict key behaviors such as lexical choice errors and improve the model's

correlation with association norms (Andrews et al., 2009). Using a similar method, Johns and Jones (2012) can predict the sensorimotor features of unseen words and fit better to WordNet feature lists than pure co-occurrence measures.

More sophisticated models forego hand-coded feature lists and incorporate techniques from the field of machine vision. Bruni et al. (2014) put forward a model that describes images as a set of visual features. They extract them for tagged images and concatenate the extracted features to the corresponding terms in a co-occurrence matrix. After performing a dimensionality reduction, the resulting word vectors correlated more highly with word similarity norms than either of them in isolation. Lazaridou et al. (2017) use a Convolutional Neural Network to extract high-level features from tagged images. Then, they train a skip-gram model to predict not only the context of the word but also its associated visual features. Using these multimodal embeddings, they can model an experiment where subjects had to learn novel terms related to novel referents with very little exposure.

LLMs, too, have begun to incorporate multimodal information to improve their performance. These models are typically trained on a combination of text and images, capable of translating one domain to the other (Ramesh et al., 2022; Rombach et al., 2022). However, to our knowledge, determining the relationship of these models specifically to links between language and sensorimotor processing is in its infancy (see Berger et al., 2022; Merkx et al., 2023; Marjieh et al., 2023, for recent attempts at this). Future work could explore how similar the linkages in these models representations of meaning are to the ones assumed to underlie human language use.

5.3. The distributional paradox

Given the challenges from embodiment and symbol grounding, we are left with a paradox. On the one hand, distributional models of semantics can represent the meaning of words in a way that seems to reflect relevant dimensions of the type of knowledge that humans have about them. Moreover, the idea behind some of them—building a space of meanings based on language exposure by tracking the statistical information present in the environment—resonates with the computations attributed to language learners by usage-based theories. On the other hand, they are at odds with some of the basic assumptions about the sensorimotor bases of human conceptual systems and their link to word meaning. They also have difficulties providing an account of reference fixing and the use of words once they have been acquired.

Recent developments in the LLM field further illustrate this tension. Particularly, the newer generations of transformer-based language models exhibit an uncanny mastery of coherent, grammatical language (Contreras Kallens et al., 2023; Piantadosi, 2023) and, through their text-completion capabilities, they are able to respond to queries from users in a conversational format. This mastery of language, and thus the quality and usefulness of the representations of meaning they build during their training, has gotten to a point where distinguishing between LLM- and human-created texts and conversational output based on surface-level traits is increasingly difficult (Sandler et al., 2024; Tang et al., 2023).

In a precursor to the embodiment and grounding challenges to distributional semantics, Searle's (1980) Chinese Room argument cautioned against attributing understanding to a

formal system that manipulates symbols without intrinsic content. However, LLMs may undermine this argument (at least in part) by posing both new and familiar challenges to the intuitions behind the argument (see Cole, 2023, for a review): what if the room operator had constructed the dictionary of symbols themselves, based on their experience with the language? What if the manipulation of symbols was expressed through long grammatical texts or dialogue difficult to distinguish from humans? What if, when prompted to do it, it could draw a semantically adequate "painting of a squirrel eating a burger" (Rombach et al., 2022, p. 6)? Here, the paradox mentioned above manifests itself in the question of what would be gained by arguing that this model does not understand, or, even more weakly, know, language and the meaning of the words it is using (see Van Dijk et al., 2023, for a related point).

In the next section, we attempt to resolve this paradox by providing a tentative answer to the more basic version of these questions, which also can be found in Nick Chater's early work. Why do distributional semantics in particular, and statistical pattern-tracking methods more generally, work so well, even though they abstract away from properties that we intuitively consider essential for meaning, cognition, and language use in humans?

6. What is distributional semantics a model of?

Our tentative hypothesis about the source of the paradox—why distributional models seem to work and not work at the same time—is that it stems from the conflation of "models of language meaning" with "models of the human conceptual system." Disentangling these two notions is crucial for understanding why distributional models can explain as much as they do while remaining fundamentally, and perhaps even inherently, limited.

6.1. Distributional models and concepts

The reason that distributional models are assumed, and thus expected, to be models of the human conceptual system is that concepts are assumed to underlie the meaning of words. Concepts are postulated mental entities—akin to packets of knowledge—that have an intentional link with the objects they represent (i.e., they are *about* them; Fodor, 1989). This includes, for example, our concept of CAT, a mental representation that we possess, that allows us to cognitively interact with them through thought, reasoning, perception, imagination, and so on (see Murphy, 2004, for a thorough review on theories of concepts).

The default interpretation of distributional models of meaning as models of the conceptual system assumes that the meaning of a word is derived from the concept to which it is mapped. In contrast, concepts, somehow,[4] have meaning by themselves (Fodor, 1989), and are responsible for a word's capacity to refer to objects in the world. Accordingly, to know the meaning of the word "cat," and thus to be able to use it effectively, necessitates having the appropriate concept CAT and a link between them. Thus, any model purporting to capture the meaning of words, including distributional models, must ultimately be a model of human concept use (e.g., Jones et al., 2015; lake & Murphy 2023).

The assumption behind this argument is that for a word to have meaning it has to be mapped onto an *isomorphic* concept—what Lupyan and Lewis (2019) critically refer to as the "words-as-mappings" view (see also Elman, 2004). This assumption has several drawbacks that seriously undermine its viability. One of them, exhaustively discussed by Fodor (1998), is that it ultimately requires an innate basis of concepts to at least kickstart the mapping process. If there is nothing to map onto in the beginning, learning cannot start from concepts; and because words are learned by mapping them onto a concept, the process cannot start with words. In contrast to Fodor (1998), we are not willing to bite that bullet. Thus, we consider this argument a reason to abandon the words-as-mappings view.

A less philosophical flaw is the separation between language and the environment that is entailed by the words-as-mapping view. This division between words and the concepts they map onto suggests that the processes of forming concepts and acquiring the labels related to them are entirely separate, either taking place in parallel or with labels being acquired after the concepts are in place. However, language is part of the environment with which humans interact, and we build our knowledge of it along with the rest of the nonlinguistic environment[5] (e.g., Tomasello, 2003). A more plausible view of the relationship between words and the rest of our mental furnishings (Prinz, 2004) is that they are not isomorphic to specific mental concepts but networked in many-to-many relationships with other words and our multifaceted experiences with the world (Elman, 2004). In this network, words themselves act as cues that activate or retrieve other related knowledge, both linguistic and nonlinguistic (Christiansen & Chater, 2022; Lupyan & Lewis, 2019). In other words, there is no sharp separation between experiences of words and the things they mean: the word "cat" is as much a part of our representation of cats as the textures, colors, and sounds associated with our experiences of these felines.

A particularly effective illustration of the explanatory power of the view that words are part of a network comprised of both linguistic and nonlinguistic experiences (Louwerse & Jeuniaux, 2010) can be found in the differences between concrete and abstract words (Montefinese, 2019). Under this view, abstract words are stronger cues for other related linguistic knowledge, through which it was probably acquired in the first place, whereas concrete words retrieve more strongly associated sensorimotor knowledge (Borghi et al., 2017; see Anderson et al., 2019, for a partial computational implementation of this idea). However, as discussed above, this type of mixture between sensorimotor and linguistic experiences, and the preferential activation of some type of content over the other, is a challenging target for models that use purely distributional cues. This is the embodied and grounded objections to distributional semantics models from a different vantage point.

This view, however, has a critical drawback. Particularly, we contend that, although it presents a powerful and viable view of the relationship between words and the conceptual system, referring to this network of knowledge cued within an individual as a word's "meaning" (Elman, 2004; Lupyan & Lewis, 2019) seems to ignore one of the key *explananda* of a usage-based account of meaning: communicative success. Indeed, communication is traditionally described as a collaborative attempt by the participants to grasp each other's intended meaning (Clark, 1996; Grice, 1957; Levinson, 2019). In the words-as-mapping view, this procedure is relatively easy to describe: the psychological concepts that the words used by

the participants are mapped onto are likely already shared. If they are not, subtle differences can get aligned during communication through mechanisms such as collaboration (Clark & Wilkes-Gibbs, 1986; Levinson, 2006), alignment (Hasson et al., 2012; Pickering & Garrod, 2004), coordination (Fusaroli et al., 2014), or prediction (Pickering & Garrod, 2013). By activating the same relatively simple concepts, people can come to use words with the same meaning (Stolk et al., 2016).

However, this story is more complicated for the words-as-cues perspective, because the information cued by a word in a specific person can be assumed to be largely idiosyncratic. It is not hard to imagine a conversation in which the activated memories of specific contexts of usage (Wojcik et al., 2022) and sensorimotor/affective states related to interactions with the referents (Lupyan & Lewis, 2019) that are associated with a word might differ radically between the participants. If these networks of "any and all information that is relevant to the use and interpretation of a word" (Elman, 2009, p. 568) for each individual are the meanings of the terms, it is challenging to explain how people come to use words with the *same meaning*, and, consequently, how they can communicate successfully.

Thus, given that language is an inherently social phenomenon, and communicative interactions are its primary setting (Beckner et al., 2009), whatever is cued by words within individuals cannot be the full story about meaning, lest we assume that meanings are not primarily for communication, but for thinking (e.g., Chomsky, 2005). This is another bullet we are not willing to bite. However, the words-as-cues view offers a more satisfactory account of the relationship between language and cognition than the words-as-mappings view. How can the conflict between this account, communication, and meaning be resolved? In what follows, we suggest that it is through cutting the link between meaning and the human conceptual system, which in turn illustrates how to interpret distributional models of meaning and their success.

6.2. Meaning, culture, and interaction

As a starting point for disentangling the tension between meaning and the cognition of individual humans, consider Strawson's (1950) distinction between "referring" and "meaning." His proposal involves a change in the locus of reference from language to the language user: reference is not something phrases or sentences *have*, but something people use them to *do*. Referring is an activity that uses language, and what can be used to refer to what is, in part, limited by their meaning. In turn, the meaning of an expression is captured by the general patterns of its use in referring. In other words, a more fruitful conception is considering the meaning of an expression to be the "general directions for its use to refer to or mention particular objects or persons" (Strawson, 1950, p. 327). Echoing Frege's (1948/1892) early insight, meaning is not reference, nor is reference necessarily mediated or enabled by it. From the point of view of the language user, learning the meaning of a word is learning how to use it to (among other things) refer in specific communicative instances with other members of a linguistic community (Wittgenstein, 1953). This notion of meaning-as-use is also at the center of Nick Chater's recent work with Morten Christiansen, providing a modern take on Wittgenstein's notion of "language games" in terms of collaborative improvisations as the foundation for the processing, acquisition, and evolution of language (Christiansen & Chater, 2022).

Our suggestion is that distributional models capture this notion of the meaning of words *as their general patterns and expectations of use*—without modeling an individual who is able to use those words (for a related argument, see Westera & Boleda, 2019). This idea has been suggested in other forms before. Landauer and Dumais (1997), for example, explicitly say that "the similarity relations between words that are extracted by LSA are based on usage" (p. 227), and not reference.[6] More recently, Baroni et al. (2014a) argued that the meanings in distributional semantics models are computational implementations of Frege's "senses," in that they can be used to induce descriptions of how objects in the world appear as opposed to being able to determine whether statements about them are true or false.[7]

We can use this idea to understand the relationship between distributional models of cognition and the cognitive systems of particular individuals. A user's previous experience with language—which we can assume to be a subsample of the space of patterns captured by contemporary distributional models—is only one of the factors that guide any particular instance of language use and which may influence their conceptual system. Other factors shaping particular usage events can include, for example, an individual's affective state (Beukeboom & Semin, 2006; Hinojosa et al., 2020), their sensorimotor history and expertise (Beilock et al., 2008; Ibáñez et al., 2023), and coordination with a conversational partner (Branigan et al., 2000; Richardson & Dale, 2005). Nonetheless, to the extent that a person's experiences with language resembles the general characteristics of the training data of a distributional model, the model will be able to capture the behavior produced by the mechanisms insofar as it is shaped by those experiences.

Intriguingly, LLMs can be seen as putting this version of meaning-as-use into actual usage in interactions with people, without assuming that they, therefore, must be viewed as *cognitive agents*: their success comes from their effectiveness at capturing the statistical patterns of a representative slice of the information that is also presumably available to humans. Our perspective explains also why these interactions—and the "human-likeness" of LLMs more generally—tend to be mediated by cultural biases in relation to, specifically, American English, given the characteristics of their training data (Atari et al., 2023; Cao et al, 2023; Johnson et al., 2022; Tao et al, 2024). Moreover, this view also predicts that distributional models of meaning can be fruitfully applied to studying differences within and between individuals in terms of their degree of reliance on (Alfred et al., 2021; Johns, 2024; Wang & Bi, 2021) and experience with (Alhama et al., 2020) public language.

Viewed like this, most of the bite of the embodiment and grounding criticisms of distributional semantics is diffused. Both criticisms point to these models' failures at capturing the cognitive and neural processes that enable humans to use language. But whether psychological or biological realism is a requirement for their success depends on *what they are assumed to be a model of*. If we assume, based on their broad functional description, that distributional models are statistical abstractions of an input space that represents the linguistic environment of humans, then they do not need also to represent the mechanism with which humans learn to interact with and use it to provide successful and effective explanations. In this sense, distributional semantics are not, in principle, required to account for specific instances of language use by humans any more than a successful model of aerodynamic phenomena must also account for the principles underlying aircraft design in aviation. But this does not mean that

they cannot inform one another. The probability that any *arbitrary* model of aerodynamics will allow us to understand how actual airplanes function is astronomically low, considering the large space of possibilities. If we found a mysterious flying machine, a relatively successful model of aerodynamics could guide our study of the workings of the machine insofar as the model is an adequate representation of the principles that inspired its design. Conversely, studying this already existing machine can help us sharpen our model of aerodynamics by trying to explain *why it works*.[8] That is why *the fact that models work in their application* can simultaneously be informative about the domain they are a model of as well as the domain they are applied to without the former also representing the latter.[9]

With this, we can return to our original paradox of why distributional models, while incapable of representing the psychological processes of language use in humans, nonetheless work as well as they do (especially the recent LLMs). We suggest that the answer is that meanings are (at least in part) the *shared environmental structures that allow us to use language socially* (Christiansen & Chater, 2022; Rączaszek-Leonardi et al., 2018; Strawson, 1950; Vygotsky, 2012; Wittgenstein, 1953). That is, we have experienced particular instances of communicative language use and learned from them using a cognitive system that is affected by those experiences. These experiences of language are not isolated, but a part of our total experience of the world, including sensorimotor and social information. Thus, language is a *shared experience* in which each individual's linguistic experience is partially overlapping with the experiences of other people in their community. It is thanks to this overlap in linguistic experience that we can use language to communicate successfully with others. And these shared experiences, at a large scale, is what an abstraction over the statistical patterns present in linguistic corpora can capture: the regularities of the linguistic environment from which humans learn to use language.

This point connects Nick Chater's early work on distributional models with his more recent work on the relationship between culture, cognition, and language. In his work with Christiansen (Christiansen & Chater, 2008, 2016, 2022), Nick Chater has helped promote the idea that language should be studied as an adaptive evolutionary system on its own right (see also Beckner et al., 2009). The properties of this system are shaped, in part, by the cognitive and communicative needs of the language users in the community that learns and uses that language. Distributional models are able to capture the patterns of meaning-through-use in this culturally shared system—the very meanings that allow speakers to communicate about whatever they need to express in the moment.

The idea of meanings as parts of a culturally evolving system provides an explanation of why distributional models can capture meaning-related phenomena in humans so well. Indeed, over the course of cultural evolution, language has adapted to fit and reflect the learning mechanisms of those who learn it (Chater & Christiansen, 2010; Christiansen & Chater, 2008, 2016). These pressures will manifest themselves through changes in the patterns of usage. However, the changes are not random: rather, they are subject to various cultural, cognitive, and environmental attractors (Beckner et al., 2009; Carr et al., 2017; Chater & Christiansen, 2022; Contreras Kallens et al., 2018). Through this evolutionary process, language will be shaped around these attractors such that individual usage events

will be roughly distributed around them. Thus, one can infer the rough shape of the shared system through the distributional properties of individual events, such as word usage.

Therefore, distributional models of meaning can be construed as models of the semantic dimension of the cultural attractors (Buskell, 2017) of language, providing an approximation of the shared cultural scaffolding (Clark, 2006) of acquisition and use (Sheya & Smith, 2019; Tomasello, 2016) based on usage events as recorded in corpora. Correspondingly, from the perspective of individual language users, distributional models describe the socially stable outcome of processes that are implemented idiosyncratically at lower levels of organization (Kelso, 1995). The idea that tracking and generalizing statistical patterns play a key role in language acquisition (e.g., Abbot-Smith & Tomasello, 2006; Saffran & Kirkham, 2018; Tomasello, 2003) further explains why *specifically distributional* models of semantics are so good at capturing these attractors, as already hypothesized in Nick Chater's early work on discovering lexical categories.

This perspective can help account for two further aspects of the fit of distributional models to human behavior. First, it can account for the differences between how closely different models can capture human meaning use, and their improvements over time: models get better at this as they approximate more closely what humans do when they engage with their speech community, the type of cues they use, what they extract from these cues, and what they use them for. This does not mean, however, that they have become better *models of humans*, but only of the type of information to which humans are sensitive (Antonello & Huth, 2023). Indeed, this predicts that some improvements of the model's mechanism to more closely mirror the distributional properties of corpora can be detrimental to their fit of the language processing of individual humans (see, e.g., Oh & Schuler, 2023). And second, the bidirectional influences between the linguistic culturally evolving system and the language users have the consequence that, over time, the cultural scaffolding of human language use will become partially grounded by proxy without needing a body or sensorimotor experiences of the environment. Accordingly, repeated studies have found that linguistic and perceptual stimuli are redundant to a surprising degree (e.g., Marjieh et al., 2023; Riordan & Jones, 2011; Willits et al., 2015).

In summary, we have argued that the distributional paradox lies in the conflation of two different dimensions of meaning. On the one hand, there are the meanings in the *community-level linguistic environment*, an abstraction of the culturally transmitted patterns of use by the members of a community that are shared between them, and which enables successful communication. On the other hand, there are the cognitive mechanisms behind our *individual ability to use language meaningfully* to perform behaviors such as referring to the world. We propose that distributional models of meaning should be seen as capturing the former rather than the latter. Furthermore, given that language is shaped by cognition (Christiansen & Chater, 2008) and acquired through statistical learning from others (Christiansen & Chater, 2022), these models work well when applied to human behavior without needing to capture the mechanisms behind human language use.

7. Conclusion

At the end of their 1992 article discussing the potential philosophical implications for theories of meaning of early connectionism, the precursors of today's LLMs, Christiansen and Chater noted that "… which philosophical challenges connectionism will generate, as well as its potential significance as a new metaphor for the mind, cannot be decided *a priori* through philosophical investigation. Rather, it is an empirical issue—only time, and the vigorous development of connectionist research techniques, will tell" (p. 247).

Since then, distributional models of language, and particularly of semantics, have come further than anyone would have expected, most vividly exemplified by the ability of the recent slate of LLMs to carry on conversations with humans. We have suggested that this has been achieved, despite the paradox stemming from their limited psychological and neural plausibility, because they capture the shared cultural scaffolding that guides the acquisition and use of language such that it can be used communicatively—and not necessarily the knowledge associated with words that enables each individual to use them meaningfully. Whereas embodied and grounded accounts of meaning attempt to describe the latter, distributional semantics models can capture the statistical structure of the former.

The extent to which distributional models can capture semantic phenomena is a testament to how humans learn language through statistical learning and use it to communicate with each other. Moreover, because language is a culturally evolving adaptive system, being well suited to be learned and useful for communication are pressures that shape its structure. This is why, in a very real sense, the meaning of a word *is* the partially stable pattern underlying its individual usage events, and thus why distributional methods are uniquely capable of capturing meaning.

In closing, explaining the success and limitations of distributional models of semantics requires abandoning both the word-as-mappings perspective and the idea that the meaning of a word is somehow captured by the cognitive or neural activity of those who use it. It also requires proponents of distributional models to abandon the notion that they are somehow models of cognition or the brain. Instead, their successes and failures are better explained as consequences of what they capture: a dimension of a culturally evolving adaptive system that guides (and is shaped by) learning and communication (Christiansen & Chater, 2008, 2016, 2022). Thus, distributional models of semantics can be seen as descriptions of meaning understood as the public, cultural infrastructure that allows individual cognitive systems to use language meaningfully. In this sense—and maybe only in this sense—interpretations of distributional models should heed Putnam's (1974) early externalist dictum: meanings are not in the head.

Notes

1 We grossly oversimply the huge literature on the philosophy of language here in the service of moving onto our main topic of discussion: distributional semantics and what it might tell about the meaning of words.

2 Or, as Firth suggests, meaning itself can be thought of as "a complex of contextual relations" (Firth, 1957, p. 6).
3 More recently, Contreras Kallens, Monaghan, and Christiansen (in preparation) have shown that phonology provides coarse constraints on the semantics of words across 200 different languages, allowing for the separation of words referring to things and actions.
4 Fodor (1998) himself later acknowledged that the huge promissory note inherent in the qualifier "somehow" has turned out to be very difficult to cash in.
5 Wittgenstein (1953, sections 26–38) makes a similar point when discussing naming as a language game.
6 This despite later arguments to the contrary by Landauer (2007).
7 Or their "mode of presentation" (Zalta, 2001).
8 Indeed, it has been argued that this interaction between aerodynamics and aviation explains a significant part of the success of both fields (Rae, 1961; see also von Kármán, 2004, particularly chapters 1 and 2).
9 See Morrison and Morgan (1999) for a discussion of how constraints on the use and manipulation of models generates knowledge.

References

Abbot-Smith, K., & Tomasello, M. (2006). Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *Linguistic Review*, *23*(3), 275–290. https://doi.org/10.1515/TLR.2006.011

Alhama, R. G., Rowland, C. F., & Kidd, E. (2020). Evaluating word embeddings for language acquisition. *Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2020)*, 38–42.

Alfred, K. L., Hillis, M. E., & Kraemer, D. J. (2021). Individual differences in the neural localization of relational networks of semantic concepts. *Journal of Cognitive Neuroscience*, *33*(3), 390-401.

Alper, M., & Averbuch-Elor, H. (2024). Kiki or bouba? Sound symbolism in vision-and-language models. *Advances in Neural Information Processing Systems*, *36*, 78347–78359.

Anderson, A. J., Binder, J. R., Fernandino, L., Humphries, C. J., Conant, L. L., Raizada, R. D. S., Lin, F., & Lalor, E. C. (2019). An integrated neural decoder of linguistic and experiential meaning. *Journal of Neuroscience*, *39*(45), 8969–8987. https://doi.org/10.1523/JNEUROSCI.2575-18.2019

Andrews, M., Frank, S., & Vigliocco, G. (2014). Reconciling embodied and distributional accounts of meaning in language. *Topics in Cognitive Science*, *6*(3), 359–370. https://doi.org/10.1111/tops.12096

Andrews, M., Vigliocco, G., & Vinson, D. P. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, *116*(3), 463–498. https://doi.org/10.1037/a0016261

Antonello, R., & Huth, A. (2023). Predictive coding or just feature discovery? An alternative account of why language models fit brain data. *Neurobiology of Language*, *5*(1), 64–79.

Asr, F. T., Willits, J. A., & Jones, M. N. (2016). Comparing predictive and co-occurrence based models of lexical semantics trained on child-directed speech. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*.

Atari, M., Xue, M. J., Park, P. S., Blasi, D., & Henrich, J. (2023). Which humans? OSF. https://doi.org/10.31234/osf.io/5b26t

Aziz-Zadeh, L., Wilson, S. M., Rizzolatti, G., & Iacoboni, M. (2006). Congruent embodied representations for visually presented actions and linguistic phrases describing actions. *Current Biology*, *16*(18), 1818–1823. https://doi.org/10.1016/j.cub.2006.07.060

Baroni, M., Bernardi, R., & Zamparelli, R. (2014a). Frege in space: A program for composition distributional semantics. In *Linguistic Issues in Language Technology, Volume 9, 2014 – Perspectives on Semantic Representations for Textual Inference*. https://aclanthology.org/2014.lilt-9.5

Baroni, M., Dinu, G., & Kruszewski, G. (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)* (pp. 238–247).

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 577–660. https://doi.org/10.1017/S0140525x99002149

Barsalou, L. W. (2016). On staying grounded and avoiding Quixotic dead ends. *Psychonomic Bulletin & Review*, *23*(4), 1122–1142. https://doi.org/10.3758/s13423-016-1028-3

Barsalou, L. W., Santos, A., Simmons, W. K., & Wilson, C. D. (2008). Language and simulation in conceptual processing. In M. de Vega, A. Glenberg, & A. Graesser (Eds.), *Symbols, embodiment, and meaning* (pp. 245–283).

Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., Holland, J., Ke, J., Larsen-Freeman, D., & Schoenemann, T. (2009). Language is a complex adaptive system: Position paper. *Language Learning*, *59*, 1–26. https://doi.org/10.1111/j.1467-9922.2009.00533.x

Beilock, S. L., Lyons, I. M., Mattarella-Micke, A., Nusbaum, H. C., & Small, S. L. (2008). Sports experience changes the neural processing of action language. *Proceedings of the national Academy of Sciences*, *105*(36), 13269-13273.

Berger, U., Stanovsky, G., Abend, O., & Frermann, L. (2022). A computational acquisition model for multimodal word categorization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 3819–3835).

Beukeboom, C. J., & Semin, G. R. (2006). How mood turns on language. *Journal of Experimental Social Psychology*, *42*(5), 553–566.

Biemiller, A., Rosenstein, M., Sparks, R., Landauer, T. K., & Foltz, P. W. (2014). Models of vocabulary acquisition: Direct tests and text-derived simulations of vocabulary growth. *Scientific Studies of Reading*, *18*(2), 130–154. https://doi.org/10.1080/10888438.2013.821992

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in Cognitive Sciences*, *15*(11), 527–536. https://doi.org/10.1016/j.tics.2011.10.001

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

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. https://doi.org/10.1162/tacl_a_00051

Bommasani, R., Davis, K., & Cardie, C. (2020). Interpreting pretrained contextualized representations via reductions to static embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4758–4781).

Borghi, A. M., Binkofski, F., Castelfranchi, C., Cimatti, F., Scorolli, C., & Tummolini, L. (2017). The challenge of abstract concepts. *Psychological Bulletin*, *143*(3), 263–292. https://doi.org/10.1037/bul0000089

Botarleanu, R. M., Dascalu, M., Watanabe, M., McNamara, D. S., & Crossley, S. A. (2021). Multilingual age of exposure. In *International Conference on Artificial Intelligence in Education* (pp. 77–87).

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition*, *75*(2), B13–B25.

Bruni, E., Tran, N. K., & Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, *49*, 1–47. https://doi.org/10.1613/jair.4135

Bullinaria, J. A., & Levy, J. P. (2013). Limiting factors for mapping corpus-based semantic representations to brain activity. *PLOS ONE*, *8*(3), e57191. https://doi.org/10.1371/journal.pone.0057191

Burgess, C., & Lund, K. (2000). The dynamics of meaning in memory. In E. Dietrich & A. B. Markman (Eds.), *Cognitive dynamics: Conceptual and representational change in humans and machines* (Vol. 13, pp. 17–56). Psychology Press.

Buskell, A. (2017). What are cultural attractors? *Biology & Philosophy*, *32*(3), 377–394. https://doi.org/10.1007/s10539-017-9570-6

Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P.S., & Sun, L. (2023). A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT. ArXiv, abs/2303.04226.

Caucheteux, C., & King, J. R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology*, *5*(1), 134.

Carr, J. W., Smith, K., Cornish, H., & Kirby, S. (2017). The cultural evolution of structured languages in an open-ended, continuous world. *Cognitive Science*, *41*(4), 892–923. https://doi.org/10.1111/cogs.12371

Chandrasekaran, D., & Mago, V. (2021). Evolution of semantic similarity—A survey. *ACM Computing Surveys (CSUR)*, *54*(2), 1–37.

Chang, T. A., & Bergen, B. K. (2022). Word acquisition in neural language models. *Transactions of the Association for Computational Linguistics*, *10*, 1–16.

Chater, N., & Christiansen, M. H. (2010). Language acquisition meets language evolution. *Cognitive Science*, *34*(7), 1131–1157. https://doi.org/10.1111/j.1551-6709.2009.01049.x

Chater, N., & Christiansen, M. H. (2022). Grammar through spontaneous order. In S. Lappin & J.-P. Bernady (Eds.), *Algebraic structures in natural language* (pp. 61–75). CRC Press.

Chomsky, N. (2005). Three factors in language design. *Linguistic Inquiry*, *36*, 1–22.

Chomsky, N. (2015). *The minimalist program* (20th Anniversary Edition). MIT Press.

Christiansen, M. H., & Chater, N. (1992). Connectionism, meaning and learning. *Connection Science*, *4*, 227–252. https://doi.org/10.1080/09540099208946617

Christiansen, M. H., & Chater, N. (1993). Symbol grounding – The emperor's new theory of meaning? In *Proceedings of the 15th Annual Cognitive Science Society Conference* (pp. 155–160). Hillsdale, NJ: Lawrence Erlbaum.

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, *31*(05), 489–509, https://doi.org/10.1017/S0140525x08004998

Christiansen, M. H., & Chater, N. (2016). *Creating language: Integrating evolution, acquisition, and processing*. MIT Press.

Christiansen, M. H., & Chater, N. (2022). *The language game: How improvisation created language and changed the world*. Basic Books.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39. https://doi.org/10.1016/0010-0277(86)90010-7

Clark, A. (2006). Language, embodiment, and the cognitive niche. *Trends in Cognitive Sciences*, *10*(8), 370–374. https://doi.org/10.1016/j.tics.2006.06.012

Clark, H. H. (1996). *Using Language*. Cambridge University Press. https://doi.org/10.1017/cbo9780511620539

Coenen, A., Reif, E., Yuan, A., Kim, B., Pearce, A., Viégas, F., & Wattenberg, M. (2019). Visualizing and measuring the geometry of BERT. *Advances in Neural Information Processing Systems*, 32.

Cole, D. (2023). The Chinese Room argument. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy (Summer 2023 Edition)*.

Contreras Kallens, P., & Dale, R. (2018). Exploratory mapping of theoretical landscapes through word use in abstracts. *Scientometrics*, *116*(3), 1641–1674. https://doi.org/10.1007/s11192-018-2811-x

Contreras Kallens, P., Dale, R., & Smaldino, P. E. (2018). Cultural evolution of categorization. *Cognitive Systems Research*, *52*, 765–774. https://doi.org/10.1016/j.cogsys.2018.08.026

Contreras Kallens, P., Kristensen-McLachlan, R. D., & Christiansen, M. H. (2023). Large language models demonstrate the potential of statistical learning in language. *Cognitive Science*, *47*(3), e13256. https://doi.org/10.1111/cogs.13256

Contreras Kallens, P., Monaghan, P., & Christiansen, M. (in preparation). *How the sounds of words can support early stages of language learning*. Unpublished manuscript, Cornell University.

Ćwiek, A., Fuchs, S., Draxler, C., Asu, E. L., Dediu, D., Hiovain, K., Kawahara, S., Koutalidis, S., Krifka, M., Lippus, P., Lupyan, G., Oh, G. E., Paul, J., Petrone, C., Ridouane, R., Reiter, S., Schümchen, N., Szalontai, Á., Ünal-Logacev, Ö., Zeller, J., … Winter, B. (2022). The bouba/kiki effect is robust across cultures and writing systems. *Philosophical Transactions of the Royal Society B*, *377*, 20200390. https://doi.org/10.1098/rstb.2020.0390

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*, 391–407.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 4171–4186).

Dingemanse, M., Blasi, D., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, iconicity and systematicity in language. *Trends in Cognitive Sciences*, *19*, 603–615. https://doi.org/10.1016/j.tics.2015.07.013

Du, J., Qi, F., & Sun, M. (2019). Using BERT for word sense disambiguation (arXiv:1909.08358). arXiv. https://doi.org/10.48550/arXiv.1909.08358

Edmiston, D. (2020). A systematic analysis of morphological content in BERT models for multiple languages (arXiv:2004.03032). arXiv. https://doi.org/10.48550/arXiv.2004.03032

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning 7*, 195–225.

Elman, J. L. (2004). An alternative view of the mental lexicon. *Trends in Cognitive Sciences*, *8*(7), 301–306. https://doi.org/10.1016/j.tics.2004.05.003

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science*, *33*(4), 547–582.

Ethayarajh, K. (2019). How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 55–65).

Evangelopoulos, N., Zhang, X., & Prybutok, V. R. (2012). Latent semantic analysis: Five methodological recommendations. *European Journal of Information Systems*, *21*(1), 70–86. https://doi.org/10.1057/ejis.2010.61

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., Tomasello, M., Mervis, C. B., & Stiles, J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, *59*(5), 1–173. https://doi.org/10.2307/1166093

Finch, S. P., & Chater, N. (1991). A hybrid approach to the automatic learning of linguistic categories. *Artificial Intelligence and Simulated Behaviour Quarterly*, *78*, 16–24.

Finch, S. P., & Chater, N. (1992). Bootstrapping syntactic categories. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society of America* (pp. 820–825). Cognitive Science Society.

Finch, S. P., & Chater, N. (1994). Distributional bootstrapping: From word class to proto-sentence. In A. Ram & K. Eiselt (Eds.), *Proceedings of the 16th Annual Meeting of the Cognitive Science Society* (pp. 301–306). Lawrence Erlbaum Associates.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930–1955. *Studies in Linguistic Analysis*.

Fodor, J. A. (1989). *Psychosemantics: The problem of meaning in the philosophy of mind*. MIT Press.

Fodor, J. A. (1998). *Concepts: Where cognitive science went wrong*. Oxford University Press.

Fodor, J. A. (2001). Language, thought and compositionality. *Royal Institute of Philosophy Supplements*, *48*, 227–242. https://doi.org/10.1017/S1358246100010808

Fourtassi, A., Scheinfeld, I., & Frank, M. (2019). The development of abstract concepts in children's early lexical networks. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (pp. 129–133).

Frege, G. (1948). Sense and Reference (M. Black, Trans.). *Philosophical Review*, *57*(3), 209–230. https://doi.org/10.2307/2181485. (Original work published in 1892).

Frermann, L., & Lapata, M. (2016). Incremental Bayesian category learning from natural language. *Cognitive Science*, *40*(6), 1333–1381. https://doi.org/10.1111/cogs.12304

Fusaroli, R., Rączaszek-Leonardi, J., & Tylén, K. (2014). Dialog as interpersonal synergy. *New Ideas in Psychology*, *32*, 147–157. https://doi.org/10.1016/j.newideapsych.2013.03.005

Gatti, D., Marelli, M., & Rinaldi, L. (2023). Out-of-vocabulary but not meaningless: Evidence for semantic-priming effects in pseudoword processing. *Journal of Experimental Psychology: General*, *152*, 851–863.

Gerz, D., Vulić, I., Hill, F., Reichart, R., & Korhonen, A. (2016). SimVerb-3500: A large-scale evaluation set of verb similarity. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 2173–2182). https://doi.org/10.18653/v1/D16-1235

Glenberg, A. M., & Robertson, D. A. (2000). Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory and Language*, *43*(3), 379–401. https://doi.org/10.1006/jmla.2000.2714

Glenberg, A. M., & Mehta, S. (2012). The limits of covariation. In *Symbols and embodiment: Debates on meaning and cognition*. Oxford University Press.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Fanda, L., Doyle, W., Friedman, D. … Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*(3), 369–380.

Grice, H. P. (1957). Meaning. *Philosophical Review*, *66*(3), 377–388.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211–244. https://doi.org/10.1037/0033-295X.114.2.211

Haber, J., & Poesio, M. (2021). Patterns of lexical ambiguity in contextualised language models (arXiv:2109.13032). arXiv. https://doi.org/10.48550/arXiv.2109.13032

Harnad, S. (1990). The symbol grounding problem. *Physica D*, *42*, 335–346.

Haslett, D. A., & Cai, Z. G. (2024). Systematic mappings of sound to meaning: A theoretical review. *Psychonomic Bulletin & Review*, *31*(2), 627–648.

Hasson, U., Ghazanfar, A. A., Galantucci, B., Garrod, S., & Keysers, C. (2012). Brain-to-brain coupling: A mechanism for creating and sharing a social world. *Trends in Cognitive Sciences*, *16*, 114–121. https://doi.org/10.1016/j.tics.2011.12.007

Hill, F., Reichart, R., & Korhonen, A. (2015). SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4), 665–695. https://doi.org/10.1162/COLI_a_00237

Hinojosa, J. A., Moreno, E. M., & Ferré, P. (2020). Affective neurolinguistics: Towards a framework for reconciling language and emotion. Language, *Cognition and Neuroscience*, *35*(7), 813-839.

Huebner, P. A., & Willits, J. A. (2018). Structured semantic knowledge can emerge automatically from predicting word sequences in child-directed speech. *Frontiers in Psychology*, *9*:133, https://doi.org/10.3389/fpsyg.2018.00133

Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458.

Ibáñez, A., Kühne, K., Miklashevsky, A., Monaco, E., Muraki, E., Ranzini, M., Speed, L. J., & Tuena, C. (2023). Ecological meanings: A consensus paper on individual differences and contextual influences in embodied language. *Journal of Cognition*, *6*(1), 59.

Imai, M., & Kita, S. (2014). The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), 20130298.

Jain, S., & Huth, A. (2018). Incorporating context into language encoding models for fMRI. *Advances in neural information processing systems*, *31*.

Johns, B. T. (2024). Determining the relativity of word meanings through the construction of individualized models of semantic memory. *Cognitive Science*, *48*(2), e13413.

Johns, B. T., & Jones, M. N. (2012). Perceptual inference through global lexical similarity. *Topics in Cognitive Science*, *4*(1), 103–120. https://doi.org/10.1111/j.1756-8765.2011.01176.x

Johnson, R. L., Pistilli, G., Menédez-González, N., Duran, L. D. D., Panai, E., Kalpokiene, J., & Bertulfo, D. J. (2022). The Ghost in the Machine has an American accent: Value conflict in GPT-3 (arXiv:2203.07785). arXiv. https://doi.org/10.48550/arXiv.2203.07785

Jones, M. N., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, *55*(4), 534–552. https://doi.org/10.1016/j.jml.2006.07.003

Jones, M. N., Willits, J., & Dennis, S. (2015). Models of semantic memory. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *The Oxford handbook of computational and mathematical psychology* (pp. 232–254). Oxford University Press.

Kelso, J. A. S. (1995). *Dynamic patterns: The self-organization of brain and behavior*. MIT Press.
Kousta, S. T., Vigliocco, G., Vinson, D. P., Andrews, M., & Del Campo, E. (2011). The representation of abstract words: Why emotion matters. *Journal of Experimental Psychology: General*, *140*(1), 14–34. https://doi.org/10.1037/a0021446
Lake, B. M., & Murphy, G. L. (2023). Word meaning in minds and machines. *Psychological Review*, *130*, 401–431. https://doi.org/10.1037/rev0000297
Landauer, T. K. (2007). LSA as a theory of meaning. In *Handbook of latent semantic analysis*. Psychology Press.
Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's Problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240.
Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259–284.
Landauer, T. K., Kireyev, K., & Panaccione, C. (2011). Word maturity: A new metric for word knowledge. *Scientific Studies of Reading*, *15*(1), 92–108. https://doi.org/10.1080/10888438.2011.536130
Langacker, R. W. (2008). *Cognitive grammar: A basic introduction*. Oxford University Press.
Lazaridou, A., Marelli, M., & Baroni, M. (2017). Multimodal word meaning induction from minimal exposure to natural text. *Cognitive Science*, *41*(S4), 677–705. https://doi.org/10.1111/cogs.12481
Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics*, *4*(1), 151–171. https://doi.org/10.1146/annurev-linguistics-030514-125254
Lenci, A., Sahlgren, M., Jeuniaux, P., Gyllensten, A. C., & Miliani, M. (2022). A comparative evaluation and analysis of three generations of distributional semantic models. *Language Resources & Evaluation 56*, 1269–1313. https://doi.org/10.1007/s10579-021-09575-z
Levinson, S. C. (2006). On the human "interactional engine". In N. J. Enfield & S. C. Levinson (Eds.), *Roots of human sociality: Culture, cognition and interaction* (pp. 39–69). Berg.
Levinson, S. C. (2019). Interactional foundations of language: The interaction engine hypothesis. In P. Hagoort (Ed.), *Human language: From genes and brain to behavior* (pp. 189–200). MIT Press.
Linzen, T., & Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics*, *7*, 195–212.
Liu, Q., Kusner, M. J., & Blunsom, P. (2020). A survey on contextual embeddings (arXiv:2003.07278). arXiv. https://doi.org/10.48550/arXiv.2003.07278
Louwerse, M. M., & Jeuniaux, P. (2010). The linguistic and embodied nature of conceptual processing. *Cognition*, *114*(1), 96–104. https://doi.org/10.1016/j.cognition.2009.09.002
Lund, K., & Burgess, C. (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
Lund, K., & Burgess, C. (1997). Modelling parsing constraints with high-dimensional context space. *Language and Cognitive Processes*, *12*(2–3), 177–210. https://doi.org/10.1080/016909697386844
Lupyan, G., & Lewis, M. (2019). From words-as-mappings to words-as-cues: The role of language in semantic knowledge. *Language, Cognition and Neuroscience*, *34*(10), 1319–1337. https://doi.org/10.1080/23273798.2017.1404114
MacWhinney, B. (2014). *The Childes Project: Tools for analyzing talk, Volume II: The database* (3rd ed.). Psychology Press. https://doi.org/10.4324/9781315805641
MacWhinney, B., & Snow, C. (1985). The child language data exchange system. *Journal of Child Language*, *12*, 271–295.
Maddison, W. P., & Maddison, D. R. (2023). Mesquite: A modular system for evolutionary analysis. Version 3.81. http://www.mesquiteproject.org
Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78. https://doi.org/10.1016/j.jml.2016.04.001
Marjieh, R., Sucholutsky, I., van Rijn, P., Jacoby, N., & Griffiths, T. L. (2023). Large language models predict human sensory judgments across six modalities (arXiv:2302.01308). arXiv. https://doi.org/10.48550/arXiv.2302.01308

Mars, M. (2022). From word embeddings to pre-trained language models: A state-of-the-art walkthrough. *Applied Sciences*, *12*(17), 8805.

Merkx, D., Scholten, S., Frank, S. L., Ernestus, M., & Scharenborg, O. (2023). Modelling human word learning and recognition using visually grounded speech. *Cognitive Computation*, *15*(1), 272–288. https://doi.org/10.1007/s12559-022-10059-7

Meteyard, L., Cuadrado, S. R., Bahrami, B., & Vigliocco, G. (2012). Coming of age: A review of embodiment and the neuroscience of semantics. *Cortex*, *48*(7), 788–804. https://doi.org/10.1016/j.cortex.2010.11.002

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space (arXiv:1301.3781). arXiv. https://doi.org/10.48550/arXiv.1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* (pp. 3111–3119).

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, *320*(5880), 1191–1195. https://doi.org/10.1126/science.1152876

Monaghan, P., Chater, N., & Christiansen, M. H. (2003). Inequality between the classes: Phonological and distributional typicality as predictors of lexical processing. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp. 810–815). Lawrence Erlbaum.

Monaghan, P., Chater, N., & Christiansen, M. H. (2005). The differential role of phonological and distributional cues in grammatical categorisation. *Cognition*, *96*, 143–182. https://doi.org/10.1016/j.cognition.2004.09.001

Monaghan, P., Christiansen, M. H., & Chater, N. (2007). The Phonological-Distributional Coherence Hypothesis: Cross-linguistic evidence in language acquisition. *Cognitive Psychology*, *55*, 259–305. https://doi.org/10.1016/j.cogpsych.2006.12.001

Montefinese, M. (2019). Semantic representation of abstract and concrete words: A minireview of neural evidence. *Journal of Neurophysiology*, *121*(5), 1585–1587. https://doi.org/10.1152/jn.00065.2019

Morrison, M., & Morgan, M. S. (1999). Models as mediating instruments. In M. S. Morgan & M. Morrison (Eds.), *Models as mediators* (pp. 10–37). Cambridge University Press.

Murphy, G. (2004). *The big book of concepts*. MIT Press.

Neelakantan, A., Xu, T., Puri, R., Radford, A., Han, J. M., Tworek, J., Yuan, Q., Tezak, N., Kim, J. W., Hallacy, C., Heidecke, J., Shyam, P., Power, B., Nekoul, T. E., Sastry, G., Krueger, G., Schnurr, D., Such, F. P., Hsu, K., … Weng, L. (2022). Text and code embeddings by contrastive pre-training (arXiv:2201.10005). arXiv. https://doi.org/10.48550/arXiv.2201.10005

Oh, B. D., & Schuler, W. (2023). Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, *11*, 336–350.

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *45*(3), 255–287. https://doi.org/10.1037/h0084295

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). https://doi.org/10.3115/v1/D14-1162

Pereira, F., Detre, G., & Botvinick, M. (2011). Generating text from functional brain images. *Frontiers in Human Neuroscience*, *5*:72, https://doi.org/10.3389/fnhum.2011.00072

Pereira, F., Gershman, S., Ritter, S., & Botvinick, M. (2016). A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive Neuropsychology*, *33*(3–4), 175–190. https://doi.org/10.1080/02643294.2016.1176907

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., Botvinick, M., & Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, *9*(1), 963.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018a). Deep contextualized word representation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 2227–2237).

Peters, M. E., Neumann, M., Zettlemoyer, L., & Yih, W. (2018b). Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 1499–1509).

Piantadosi, S. (2023). Modern language models refute Chomsky's approach to language. Lingbuzz preprint, Lingbuzz.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(2), 169–190.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*(4), 329–347. https://doi.org/10.1017/S0140525x12001495

Plato. (1999). *Cratylus* (B. Jowett, Trans). Project Gutenberg. https://www.gutenberg.org/files/1616/1616-h/1616-h.htm

Portelance, E., Duan, Y., Frank, M. C., & Lupyan, G. (2023). Predicting Age of Acquisition for children's early vocabulary in five languages using language model surprisal. *Cognitive Science*, *47*(9), e13334.

Potamias, R. A., Siolas, G., & Stafylopatis, A. (2020). A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications*, *32*, 17309–17320. https://doi.org/10.1007/s00521-020-05102-3

Prinz, J. J. (2004). *Furnishing the mind: Concepts and their perceptual basis*. MIT Press.

Putnam, H. (1974). Meaning and reference. *Journal of Philosophy*, *70*(19), 699–711. https://doi.org/10.2307/2025079

Rączaszek-Leonardi, J., Nomikou, I., Rohlfing, K. J., & Deacon, T. W. (2018). Language development from an ecological perspective: Ecologically valid ways to abstract symbols. *Ecological Psychology*, *30*(1), 39–73. https://doi.org/10.1080/10407413.2017.1410387

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*. OpenAI.

Rae, J. B. (1961). Science and engineering in the history of aviation. *Technology and Culture*, *2*(4), 391–399.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(140), 1–67.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with CLIP latents (arXiv:2204.06125). arXiv. https://doi.org/10.48550/arXiv.2204.06125

Redington, M., & Chater, N. (1998). Connectionist and statistical approaches to language acquisition: A distributional perspective. *Language and Cognitive Processes*, *13*(2–3), 129–191.

Redington, M., Chater, N., & Finch, S. (1993). Distributional information and the acquisition of linguistic categories: A statistical approach. In *Proceedings of the 15th Annual Meeting of the Cognitive Science Society* (pp. 848–853). Lawrence Erlbaum Associates Inc.

Redington, M., Chater, N., & Finch, S. (1998). The potential contribution of distributional information to early syntactic category acquisition. *Cognitive Science*, *22*, 425–469.

Redington, M., Chater, N., Huang, C., Chang, L., Finch, S., & Chen, K. (1995). The universality of simple distributional methods: Identifying syntactic categories in Chinese. In *Proceedings of the Cognitive Science of Natural Language Processing*. Dublin City University.

Richardson, D. C., & Dale, R. (2005). Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science*, *29*(6), 1045–1060.

Riordan, B., & Jones, M. N. (2011). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, *3*(2), 303–345. https://doi.org/10.1111/j.1756-8765.2010.01111.x

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10684–10695).

Saffran, J. R., & Kirkham, N. Z. (2018). Infant statistical learning. *Annual Review of Psychology*, *69*(1), 181–203. https://doi.org/10.1146/annurev-psych-122216-011805

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620. https://doi.org/10.1145/361219.361220

Sandler, M., Choung, H., Ross, A., & David, P. (2024). A linguistic comparison between human and ChatGPT-generated conversations (arXiv:2401.16587). arXiv. https://doi.org/10.48550/arXiv.2401.16587

Sathvik Nair, M. S., & Meylan, S. (2020). Contextualized word embeddings encode aspects of human-like word sense knowledge. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon* (pp. 129—141).

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424. https://doi.org/10.1017/S0140525x00005756

Sheya, A., & Smith, L. (2019). Development weaves brains, bodies and environments into cognition. *Language, Cognition and Neuroscience*, *34*(10), 1266–1273. https://doi.org/10.1080/23273798.2018.1489065

Steyvers, M. (2010). Combining feature norms and text data with topic models. *Acta Psychologica*, *133*(3), 234–243. https://doi.org/10.1016/j.actpsy.2009.10.010

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (p. 15). Routledge.

Stolk, A., Verhagen, L., & Toni, I. (2016). Conceptual alignment: How brains achieve mutual understanding. *Trends in Cognitive Sciences*, *20*(3), 180–191. https://doi.org/10.1016/j.tics.2015.11.007

Strawson, P. F. (1950). On referring. *Mind*, *59*(235), 320–344.

Sun, J., Wang, S., Zhang, J., & Zong, C. (2020). Neural encoding and decoding with distributed sentence representations. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(2), 589–603.

Tang, R., Chuang, Y.-N., & Hu, X. (2023). The science of detecting LLM-generated texts (arXiv:2303.07205). arXiv. https://doi.org/10.48550/arXiv.2303.07205

Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large language models (arXiv:2311.14096). arXiv. https://arxiv.org/abs/2311.14096

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

Tomasello, M. (2016). Cultural learning redux. *Child Development*, *87*(3), 643–653. https://doi.org/10.1111/cdev.12499

van Dijk, B. M. A., Kouwenhoven, T., Spruit, M. R., & van Duijn, M. J. (2023). Large language models: The need for nuance in current debates and a pragmatic perspective on understanding (arXiv:2310.19671). arXiv. https://doi.org/10.48550/arXiv.2310.19671

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. ukasz, & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. *30*).

Vigliocco, G., Vinson, D. P., Druks, J., Barber, H., & Cappa, S. F. (2011). Nouns and verbs in the brain: A review of behavioural, electrophysiological, neuropsychological and imaging studies. *Neuroscience & Biobehavioral Reviews*, *35*(3), 407–426. https://doi.org/10.1016/j.neubiorev.2010.04.007

Von Kármán, T. (2004). *Aerodynamics: Selected topics in the light of their historical development*. Courier Corporation.

Vulić, I., Ponti, E. M., Litschko, R., Glavaš, G., & Korhonen, A. (2020). Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 7222–7240).

Vygotsky, L. S. (2012). *Thought and language*. MIT Press.

Wang, X., & Bi, Y. (2021). Idiosyncratic Tower of Babel: Individual differences in word-meaning representation increase as word abstractness increases. *Psychological Science*, *32*(10), 1617–1635. https://doi.org/10.1177/09567976211003877

Wang, B., Wang, A., Chen, F., Wang, Y., & Kuo, C. C. J. (2019). Evaluating word embedding models: Methods and experimental results. *APSIPA Transactions on Signal and Information Processing*, *8*, e19.

Wang, W., Vong, W. K., Kim, N., & Lake, B. M. (2023). Finding structure in one child's linguistic experience. *Cognitive Science*, *47*(6), e13305.

Westera, M., & Boleda, G. (2019). Don't blame distributional semantics if it can't do entailment (arXiv:1905.07356). arXiv. https://doi.org/10.48550/arXiv.1905.07356

Wiedemann, G., Remus, S., Chawla, A., & Biemann, C. (2019). Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings (arXiv:1909.10430). arXiv. https://doi.org/10.48550/arXiv.1909.10430

Willits, J. A., Amato, M. S., & MacDonald, M. C. (2015). Language knowledge and event knowledge in language use. *Cognitive Psychology*, *78*, 1–27. https://doi.org/10.1016/j.cogpsych.2015.02.002

Wittgenstein, L. (1953). *Philosophical investigations*. Blackwell.

Wojcik, E. H., Zettersten, M., & Benitez, V. L. (2022). The map trap: Why and how word learning research should move beyond mapping. *Wiley Interdisciplinary Reviews: Cognitive Science*, *13*(4), e1596.

Zalta, E. N. (2001). Fregean senses, modes of presentation, and concepts. *Philosophical Perspectives*, *15*, 335–359.

Zwaan, R. A. (2003). The immersed experiencer: Toward an embodied theory of language comprehension. *Psychology of Learning and Motivation*, *44*, 35–62. https://doi.org/10.1016/S0079-7421(03)44002-4

Zwaan, R. A. (2014). Embodiment and language comprehension: Reframing the discussion. *Trends in Cognitive Sciences*, *18*(5), 229–234. https://doi.org/10.1016/j.tics.2014.02.008