ANNUAL REVIEWS

Annual Review of Linguistics

The Probabilistic Turn in Semantics and Pragmatics

Katrin Erk

Department of Linguistics, University of Texas at Austin, Austin, Texas, USA;
email: katrin.erk@utexas.edu

Annu. Rev. Linguist. 2022. 8:101–21

First published as a Review in Advance on October 5, 2021

The *Annual Review of Linguistics* is online at linguistics.annualreviews.org

https://doi.org/10.1146/annurev-linguistics-031120-015515

Keywords

semantics, pragmatics, lexical semantics, probabilities, Bayesian models, machine learning

Abstract

This article provides an overview of graded and probabilistic approaches in semantics and pragmatics. These approaches share a common set of core research goals: (*a*) a concern with phenomena that are best described as graded, including a vast lexicon of words whose meanings adapt flexibly to the contexts in which they are used, as well as reasoning under uncertainty about interlocutors, their goals, and their strategies; (*b*) the need to show that representations are learnable, i.e., that a listener can learn semantic representations and pragmatic reasoning from data; (*c*) an emphasis on empirical evaluation against experimental data or corpus data at scale; and (*d*) scaling up to the full size of the lexicon. The methods used are sometimes explicitly probabilistic and sometimes not. Previously, there were assumed to be clear boundaries among probabilistic frameworks, classifiers in machine learning, and distributional approaches, but these boundaries have been blurred. Frameworks in semantics and pragmatics use all three of these, sometimes in combination, to address the four core research questions above.

1. INTRODUCTION

There is a trend in semantics and pragmatics that could awkwardly be described as probabilistic-ish. Not all of the works that I consider part of this trend are explicitly probabilistic, nor do they all share a single common technique. Rather, what centrally connects them is a set of common research goals—goals that are generally associated with probabilistic formalisms.

1.1. Gradience in Phenomena

Kintsch (2007) describes word meanings as "fluid and flexible." This is obvious in the case of vague adjectives like *tall*, where there is no clear boundary between tall and nontall. But most lexical items have multiple meanings, often closely related ones; Zeevat et al. (2017) count 78 senses for the verb *tall*. Probabilistic and graded approaches can then be used to describe similarities between meanings, as well as degrees of influence of context on sense choice. Cooper et al. (2015, p. 2) strongly advocate for semantic frameworks that are graded, and write of semantic frameworks that assume categorical distinctions: "they cannot represent the gradience of semantic properties that is pervasive in speakers' judgements concerning truth, predication, and meaning relations."

Similarly, Franke & Jäger (2016, p. 9) write that pragmatics "is a fuzzy and gooey affair." The aims, beliefs, and preferences of speakers and listeners are often best described in graded or probabilistic terms.

1.2. Learning

How do speakers learn to understand and produce language? Are the semantic formalisms that we use learnable? Cooper et al. (2015, p. 2) write:

> There is a fair amount of evidence indicating that language acquisition in general crucially relies on probabilistic learning. . . . It is not clear how a reasonable account of semantic learning could be constructed on the basis of the categorical type systems that either classical or revised semantic theories assume.[1]

1.3. Empirical Evaluation

Probabilistic (and -ish) approaches in semantics and pragmatics emphasize evaluating formalisms empirically, against data collected from human participants. In contrast, armchair analyses historically did not prioritize large data coverage. In pragmatics, empirical analysis often takes the form of reference games and similar experimental settings that probe human behavior for a particular pragmatic phenomenon. In lexical semantics, empirical evaluation typically means large-scale evaluation against existing data sets of word-similarity judgments, paraphrases, or entailments.

1.4. Scaling Up

Baroni et al. (2014, p. 245) write that "the problem of lexical semantics is primarily a problem of size." There are incredibly many words in the lexicon, which again can have multiple meanings, most of them not regular but idiosyncratic. Again according to Baroni et al. (2014, p. 241):

[1]Most attention has been on learning the lexicon, and this topic is also what I focus on in this review. But Liang & Potts (2015) draw a connection to the literature on "semantic parsing" in computational linguistics: the use of classifiers to learn from data how to map text to logical form. Their point is that even logical form is learnable from data; the question is simply what the right data to learn from will be.

"Statistical semantics has addressed the issue of the vastness of word meaning by proposing methods to harvest meaning automatically from large collections of text (corpora)."

In terms of techniques, some approaches use an explicitly probabilistic framework, which is prevalent in, for example, empirical pragmatics. Other approaches use machine learning techniques that used to be viewed as distinct from probabilistic frameworks, but (somewhat appropriately) the boundary has recently become uncertain and blurry. Therefore, I use "graded and probabilistic" as an umbrella term for the techniques that I discuss in this review.

The use of probabilities of in semantics and pragmatics is not a recent idea. For example, Kamp (1975) proposes a probabilistic formulation for the semantics of vague adjectives, Cohen (1999) for generic expressions, and McCready & Ogata (2007) for evidentials. In pragmatics, Parikh (2000) and Merin (1999) use probabilistic frameworks. These earlier approaches differ starkly from the more recent ones on which I focus in this article; this difference is particularly apparent in comparison with the overview article by Cohen (2003). The earlier approaches specified abstract properties of probability measures without committing to any particular probabilistic model. Now, probabilistic models are usually specified in detail and implemented in practice. Regarding the earlier approaches, Cohen (2003, p. 377) writes: "Probabilistic notions do not change the fundamentals of semantic theory, but interact with them in specific cases in order to solve a specific problem." Recent approaches assign a more central role to probabilities and aim to cover many phenomena within a single framework.

2. PROBABILISTIC FRAMEWORKS FOR SEMANTICS AND PRAGMATICS

The basics of probability distributions are quickly told. A random variable is a variable whose value depends on some random phenomenon, such as the outcome of a coin flip. The values that the random variable can take are the outcomes of a random experiment. For a coin, the value can be heads or tails. A probability distribution assigns probabilities to the different outcomes of an experiment. Probabilities are nonnegative numbers that are never greater than one. For a fair coin, either outcome has a probability of 0.5.

There are many different perspectives on probabilities, and many uses of probabilities. In the rest of this section I focus on perspectives that have been influential in semantics and pragmatics.

2.1. Bayesian Probability

There are different views of what probabilities actually are (discussed in, e.g., Hacking 2001). In Bayesian frameworks, probabilities are viewed as degrees of subjective belief, or as reasonable expectations. For example, say I know that my friend has a pet newt called Fluffy, but I don't know if Fluffy is an orange newt or a green newt. My belief state can be described through two hypotheses: h_1 = "Fluffy is an orange newt" and h_2 = "Fluffy is a green newt." I ascribe probabilities to these two hypotheses that reflect my degree of belief. Say I ascribe a probability of 0.5 to each of them. When I encounter some new evidence e, it may change my belief state by changing my degree of belief in h_1 and h_2. In a Bayesian framework, this belief update is modeled through Bayes's rule, which in this case takes the following form:

$$p(h_1|e) = \frac{p(e|h_1)p(h_1)}{p(e|h_1)p(h_1) + p(e|h_2)p(h_2)}.$$

My posterior belief in hypothesis h_1, after seeing the evidence, is $p(h_1|e)$. Bayes's rule describes it as depending on my prior belief in hypothesis h_1, $p(h_1)$, along with the likelihood $p(e|h_1)$: If h_1 were actually the correct hypothesis, how likely would I have been to see the evidence? The

denominator normalizes the combination of these two probabilities to again be a probability. The likelihood $p(h_1|e)$ is a conditional probability, the probability of h_1 given e, defined in general as $p(A|B) = p(A \wedge B)/p(B)$.

Let us do a worked example of belief update. Say the evidence is that I see my friend with a Band-Aid on her finger and she says that Fluffy bit her. Newts can be fierce creatures: I know for a fact that any orange newt has a probability $p = 0.2$ of being a biter, and for a green newt it is even higher, at $p = 0.6$. How does the evidence affect my belief state? The probability $p(e|h_1)$ of Fluffy biting my friend if he is an orange newt is 0.2, while $p(e|h_2)$ is 0.6. Following Bayes's rule, I update my belief to a posterior of $p(h_1|e) = (0.2 \times 0.5)/[(0.2 \times 0.5) + (0.6 \times 0.5)] = 0.25$. I do not completely make up my mind; I retain some uncertainty about h_1 and h_2. But I now consider it much more likely that Fluffy is a green newt.

Bayes's rule integrates two different influences: the prior belief in a hypothesis and the likelihood, which is the probability of the evidence under the hypothesis. Both probabilities influence the outcome. The stronger the evidence is, the more radically I will change my belief: If orange newts are biters with $p = 0.01$, then my belief in hypothesis h_1 after seeing my friend with a Band-Aid would go down to $p = 0.02$. And the stronger the prior is, the more hesitant I will be to change my belief: If I was almost certain that Fluffy is an orange newt ($p = 0.9$), but then I learn that he has bitten my friend, my posterior belief in h_1 would go to $p = 0.38$. By integrating e into my belief state, I have done inductive inference, from a piece of observed data (e) to its possible underlying causes. Bayes's rule can be applied repeatedly to model an agent who updates their belief state every time a new piece of evidence comes in.

2.2. Graphical Models

Complex joint distributions of many random variables, say, $p(A \wedge B \wedge C \wedge D \wedge E \wedge F)$, can be hard to estimate. But often there is some structure among the random variables: There are dependencies among some of them, but not all. Such structures are formalized in graphical models, graphs where the nodes are random variables and the edges are dependencies (Pearl 1988, Koller & Friedman 2009). **Figure 1** shows a standard example: If the sprinkler was on, the street will likely be wet. If it has rained, the street will also likely be wet—and the sprinklers likely did not run, if their sensors worked correctly. So the relevant probabilities are p(rain) and the conditional probabilities are p(sprinkler|rain) and p(street wet|sprinkler,rain). Graphical models can be used to reason over joint distributions, for example, from the observation that the street is wet to likely causes.

In the Fluffy example above, as well as in the sprinkler example, the probability distribution associated with each node is a Bernoulli distribution, which describes an experiment with a yes-or-no

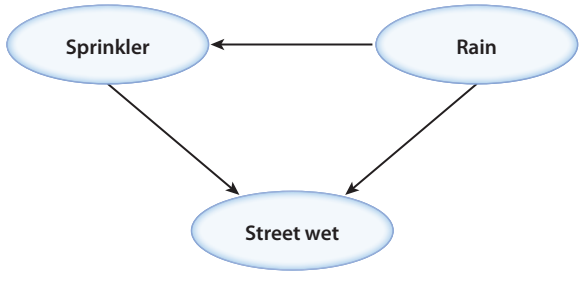

Figure 1

A simple probabilistic graphical model that represents dependencies among random variables in a joint distribution.

outcome (e.g., the street either is or is not wet). A powerful tool in probabilistic approaches is that there are many "off-the-shelf" probability distributions to characterize different kinds of random variables. Heights and weights of people and animals are well described by Gaussian distributions (bell curves). The outcome of a die roll can be described through a categorical distribution, for example, ascribing a probability of one-sixth to each outcome for a fair die.

2.3. Inference and Learning

In applying Bayes's rule, the agent performs Bayesian inference to update their beliefs; more specifically, they adjust the probabilities of the hypotheses that they consider, based on data. To put it more succinctly: They learn from observations. In the Fluffy example above, this inference is straightforward, but in larger graphical models, it can be complex.

Many standard methods for probabilistic inference exist. Sampling methods draw random values in each random variable in the graphical model. Variational methods approximate random variables by making additional assumptions about their probability distributions. Another option is to use classifiers that are general function approximators, as described next.

2.4. Probability Functions and Classifiers

Machine learning models use data to learn how to perform some task. Neural models, currently the most widely used form of machine learning model, used to be considered a framework distinct from and incompatible with Bayesian models, but the boundary has been blurred on both theoretical and practical levels (Mcclelland 2013). The most common neural models are classifiers: Given a data point, they categorize it into one of n possible categories (or assign weights to the different categories). For example, a classifier to distinguish green from orange newts would have two output classes. This simplest case represents a logistic regression model (**Figure 2*a***). It classifies data points into binary categories, zero or one, encoding orange or green (y in **Figure 2**). Each data point is described through a sequence of numeric features. These are x_1 and x_2 in **Figure 2**—say, x_1 for whether the newt bites and x_2 for whether it likes to dance. The classifier learns weights w_1 and w_2 for the features along with a bias b, a general propensity for answering one rather than zero.

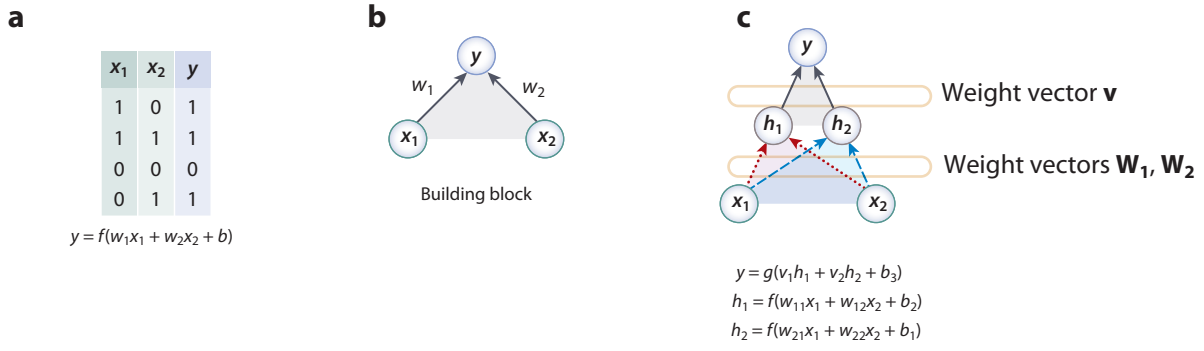

Figure 2

Neural models are universal function approximators and can also approximate probability functions. (*a*) The formula for logistic regression, with made-up sample data. Prediction y is based on a weighted sum of the inputs x_1 and x_2. w_1 and w_2 are weights, and b is the bias. (*b*) The logistic regression model can also be drawn as a graph, where weights w_1 and w_2 are now labels on the edges. For readability, we omit the bias, which is another node with an edge into y. (*c*) More-complex neural networks duplicate building blocks like the logistic regression network in panel *b*. This model has two logistic regression building blocks (*arrows*) on the bottom and another on top. The vector $\mathbf{W}_1$ of weights is $\langle w_{11}, w_{12} \rangle$, and likewise for $\mathbf{W}_2$ and $\mathbf{v}$. Again, we omit the bias.

It learns by starting with arbitrary weights and iteratively adjusting them to best fit a set of training data points (**Figure 2*a***). The classifier's response to a data point $\langle x_1, x_2 \rangle$ is then $y = f(w_1 x_1 + w_2 x_2 + b)$, a weighted sum of the values with some nonlinear transformation, such as scaling the resulting weight to be a value between zero and one. We can also draw this logistic regression model as a graph (**Figure 2*b***), where the weights w_1 and w_2 are now labels on the edges. More-complex neural models duplicate building blocks like this logistic regression network (**Figure 2*c***). The model shown in **Figure 2*c*** has two logistic regression building blocks on the bottom and another one on the top. It computes a weighted sum of the input features, transformed by f, and stores it in h_1. It also computes another weighted sum of the input features, with other weights, and stores its f transformation in h_2. The final output y is then a g-transformed weighted sum of h_1 and h_2. The model can flexibly learn combinations and transformations of the inputs in whichever way yields the best classification results. Such a model can learn to approximate any function $f(\mathbf{x}) = y$ based on examples of x–y pairs. The connection to probabilistic approaches is as follows: A probability function is a function, too, and a neural model can learn to approximate it. This is particularly important for complex probabilistic models.

2.5. Probabilistic Models in Semantics and Pragmatics

The different models described above are a good match for the research questions posed in Section 1. Bayesian models can model the belief state of a cognizer as a distribution over hypotheses, and they can model one speaker as reasoning about another, as probabilistic inference over their likely belief state. Bayesian models and more-general classifiers can learn from observed data, so they can be used to model a speaker who learns from experience. These models can be evaluated empirically by testing how well their inferences, or their guessed y labels, coincide with observations from human participants. Graphical models can be used to describe complex constellations of interacting constraints with many sources of uncertainty. This ability of graphical models to describe interacting constraints is especially important for lexical meaning in context, where the meaning of each word depends on the meanings of all other words. The structure of a graphical model can express hypotheses about a linguistic phenomenon: Which random variables interact directly, and where do we assume independence? For example, concerning word meaning in context, do we assume that the meanings of all words in a sentence directly interact, or do we assume dependencies only between direct syntactic or semantic neighbors, that is, words linked by a dependency or a semantic role?

Probabilistic programming languages like Church (Goodman et al. 2008) and WebPPL (Goodman & Stuhlmüller 2014) have probabilistic operations that mimic, for example, the random outcome of a coin flip or a die roll. They allow for a direct implementation of graphical models. They have proven useful in experimental pragmatics and have also been used in probabilistic sentence understanding models, in particular by Goodman & Lassiter (2015), Bernardy et al. (2019b), and Erk & Herbelot (2021), because they provide readable formalizations in a similar way to logic, except that they are probabilistic.

When probability functions are simply functions to be approximated, complex graphical models can be reformulated as classifiers (e.g., Emerson 2020a, White et al. 2020) for more straightforward training and better cognitive plausibility. I return to this point in Section 4.2, below.

3. BAYESIAN REASONING FOR EMPIRICAL PRAGMATICS

Franke & Jäger (2016) argue that "Bayes' rule is probably important for pragmatics" (as per the title of their paper). One main reason, they write, is that Bayesian frameworks can explain many

pragmatic effects simply through a speaker and a listener rationally and near-optimally reasoning over each other, without additional assumptions about the structure of meaning representations. The basis for this mutual reasoning between speaker and listener is that each assumes that the other abides by Grice's (1975) cooperative principle. Here, rationality means that humans are rational probabilistic reasoners (Oaksford & Chater 2007). Speakers and listeners follow rules in their reasoning, but these are probabilistic rather than logical rules, and their beliefs are subjective beliefs described in terms of probabilities.

3.1. Reference Games

Another advantage of Bayesian approaches (Franke & Jäger 2016) is that they can be evaluated against empirical data from experiments. An experimental framework that has been particularly fruitful for empirical pragmatics is the reference game (Benz & Stevens 2018). In this experimental setup, both a speaker and a listener see the same group of objects. The speaker's task is to use an utterance to point out a particular object to the listener. The speaker has two possibly competing goals: Their utterance should let the listener pick out the right object (it should be informative), and it should not be unnecessarily long-winded and laborious (it should be low cost).

3.2. Mutual Reasoning Between Speaker and Listener

As an example of mutual reasoning between speaker and listener, consider Rational Speech Acts theory (RSA) (Frank & Goodman 2012, Goodman & Frank 2016), a prominent recent framework in empirical pragmatics. In RSA, a speaker and a listener reason over each other's belief states, in several layers. Say we have a reference game. In the bottom layer, the literal listener, Listener 0, uses Bayes's rule to integrate their prior probability $p(o)$ of different objects o—how probable is it that anyone would refer to them?—with the likelihood $p(u|o)$ of an utterance u for object o, which is equal to one if u is true of o and zero otherwise. This integration of prior with likelihood yields their posterior $p_{\text{Listener0}}(o|u)$. In the next layer, a pragmatic speaker, Speaker 1, computes $p_{\text{Speaker1}}(u|o)$: They decide on an utterance u given the object o they have decided to refer to. They do so by weighing the utility or informativity of u—how likely it is to make the listener pick out the right object—against the cost of the utterance, where longer and more involved utterances would be more costly. The pragmatic speaker computes its probability using the softmax function, a variant of Luce's choice axiom that normalizes weights into probabilities. (Probabilities are proportional to the exponential of the input weights, so higher weights are strongly emphasized.) A pragmatic listener, Listener 1, then computes their posterior $p_{\text{Listener1}}(o|u)$, again using Bayes's rule, but they now reason over the pragmatic speaker: The prior $p(o)$ is the same, but the likelihood is now $p_{\text{Speaker1}}(u|o)$, the probability with which the pragmatic speaker prefers to use utterance u to refer to o.

Beyond the reference game, the same formulation can also be used for a speaker to refer to a world w via utterance u: The literal listener combines $p(w)$ with $p(u|w)$, which is now equal to one if u is true of world w and false otherwise. Speaker 1 can compute $p_{\text{Speaker1}}(u|w)$, their preference for utterance u given that the world they want to convey is w, and so on.

3.3. Many Variants Within a Common Framework

Bayesian formulations in pragmatics allow for many variations that implement different assumptions about the goals and preferences of speakers and listeners. Qing & Franke (2015) explore variants of RSA that take the salience of objects into account and that optimize with a view to the listener's action (the single object they point to in the reference game) rather than the listener's

belief (which is distribution over objects). Degen et al. (2020) turn the likelihood $p(u|w)$, the probability that utterance u is true in world w, into an actual probability: For the original literal listener, evaluation of truth in a world is binary, and $p(u|w)$ is either one or zero. According to Degen et al., an object can be considered to be blue to varying degrees.[2] McMahan & Stone (2015) model a speaker whose goal is conformity rather than informativeness. They give a pragmatic account of the interpretation of color terms, in which speakers choose terms according to their availability. For example, 'pink' would be more available (i.e., more widely used) than 'fuchsia' and hence preferred.

Bayesian formulations can also be used to formalize social aspects of pragmatic reasoning. Noble et al. (2020) use a Bayesian model to formalize reasoning over personae, ideological stances that interlocutors take in a dialogue.

Complex Bayesian models, with a pragmatic speaker and listener reasoning about each other, can be costly to compute—so is it plausible to assume that humans do this computation every time? White et al. (2020) exploit the ability of classifiers to approximate probability functions, even complex ones. They train a classifier to approximate the pragmatic speaker's probability in one step. The argument is that speakers do use mutual Bayesian reasoning but that they learn to do so implicitly, having memorized frequently needed reasoning steps.

Many different formalizations are being used in empirical pragmatics. But they are all variants of a common framework that uses Bayesian formulations to describe the mutual reasoning of speaker and listener about each other, and they agree in their focus on empirical data.

4. PROBABILISTIC AND GRADED FEATURES IN THE LEXICON

Feature-based representations are among the most widely used mechanisms for representing lexical knowledge, either as simple feature lists or structured into attribute–value matrices. More recently, features have been endowed with probabilities and gradience in order to capture the "fluid and flexible" nature of word meanings (Kintsch 2007). Most words are polysemous; they have multiple meanings that may be related to one another to a greater or lesser degree. In addition, context can strongly affect the meaning of a word, to such an extent that it is sometimes unclear what should count as a sense and what as a context effect (Tuggy 1993, Falkum & Vicente 2015).

There is a clear common aim of approaches that combine lexical features with probabilities and gradience: to give an account of word meaning in context, to model phenomena like vagueness that involve degrees, and to show how lexical knowledge can be acquired by a learner. However, at present there is no consensus on which representations or methods to use. I describe some of the trends and main questions in the following subsections.

4.1. Probabilistic Features

In this section, I look at three topics that work with probabilistic features has focused on: how a cognizer could learn such features, how they can be used to describe the meaning of vague words and expressions, and how they can be used to describe meaning in context.

4.1.1. Learning. In probabilistic feature approaches to the lexicon, the value of a feature is described as a probability distribution over values. For the simplest case of a binary feature, this is

[2] It is not clear to me whether this probability reflects a cognizer's subjective belief in an object's blueness, or perhaps instead some frequency with which the cognizer has seen speakers refer to such a hue as blue, along the lines of Emerson (2018).

a Bernoulli distribution: How likely is the agent of *eating* to be animate? (This example assumes a structured representation of *eat* that includes information on role fillers.) For the height of adult women, possible values are better described through a Gaussian distribution (a bell curve). Such probabilities can be learned from data. In the simplest case, the probability of, say, eaters being animate can be estimated as the relative frequency of co-occurrence in observed data. This ability to learn from data has been used as a proof of concept that linguistic information, including lexical information, is learnable, especially by Cooper et al. (2015), Zeevat (2015), Zeevat et al. (2017), and Herbelot & Copestake (2021).

There are many other options to mathematically model learning, drawing on Bayesian learning or other machine learning techniques. Larsson (2013), like Cooper (2021), links linguistic meaning to perception. When we perceive objects in the world, we perceive them as members of a category; for example, we may perceive an object as a tree. In doing so, we categorize the object as matching the linguistic description 'tree'. This categorization can be formalized as classification in the machine learning sense: Given perceptual features of the object, the classifier decides whether the label 'tree' applies to the object or not. Larsson uses a simple classifier similar to the one in **Figure 2*b***. The weights (the *w*s) of such a classifier can be learned from observations, called training data in machine learning. Observations here are objects with labels, for example, an object labeled as a 'tree'. The model tests itself on whether it would also call the object a 'tree', and iteratively adjusts its weights so that it will be more inclined to call the object a 'tree'. Larsson uses the idea of perception as classification to address a long-standing and tricky question: If meanings are private to each cognizer, how can two people ever successfully communicate? (For a discussion of this problem, see, for example, Pelletier 2017.) Larsson's answer is that individual agents learn a "take" on the world and coordinate their takes through linguistic interaction. For a simple game of learning the terms 'left' and 'right', Larsson shows how two speakers can coordinate their classifiers by observing each other's categorization labels and adjusting their classifier weights accordingly.

4.1.2. Vagueness. Vague adjectives like *tall* do not make a clear-cut distinction—there is no clear boundary between tall and nontall (Solt 2015). Furthermore, a *tall basketball player* is of a quite different height than a *tall toddler*. Sutton (2015) uses probabilistic features to formalize the meaning of vague adjectives as transformations on the probabilistic representations of their head nouns. Say the representation of *woman* has a *height* feature that is a distribution over possible heights. Then the representation of *tall* is an operation that shifts the value distribution of the *height* feature in its head noun. If the listener hears *Mary is tall*, this shifts their mental representation of Mary's possible heights, concentrating probability mass on higher heights and reducing their uncertainty about Mary's height. Fernández & Larsson (2014) propose an extension of the classifier approach of Larsson (2013) to handle vague adjectives. For vagueness there are probabilistic approaches with either a focus on lexical semantics or a focus on pragmatics; Lassiter & Goodman (2017) propose an RSA approach to vagueness.

Probabilistic approaches can also be used for vague adjectives that cannot easily be described as varying along a single dimension. McMahan & Stone (2015) formalize color words as coherent regions in a space of hues, where each word is characterized by a core area and a probabilistic boundary. Their approach is interesting from a perspective of learning as well as vagueness: They show that the parameters of their Bayesian model can be learned from data, in particular from an existing data set of color patches. Monroe et al. (2017) model color-naming preferences based on RSA, with separate machine learning models for the speaker and the listener.

4.1.3. Meaning in context. Zeevat et al. (2017) use probabilistic feature representations to handle polysemy and disambiguation. Instead of having different representations for different senses, they assume an overspecified feature representation: a single feature representation across all different uses of a word that may contain incompatible features. When a word appears in context, only the contextually appropriate features from the overspecified representation are retained, where appropriateness is formalized through co-occurrence probabilities that link features of a word to features of its context. Schuster et al. (2020) add probabilities between features of the same word in order to encode co-occurrence of features; for example, birds that swim tend to have webbed feet.

So far, I have concentrated on probabilistic features that are individually interpretable, such as the height of a person. I now turn to a strand of work that uses opaque features and has been highly influential in probabilistic lexical approaches: distributional models.

4.2. Distributional Models: Characterizing Word Meaning Through Observed Contexts

Words that are similar in meaning tend to appear in similar textual contexts. Distributional models use this observation to derive feature representations of word meaning automatically from corpus data. They characterize the meaning of a word through the contexts in which it has been observed in a corpus. In the simplest case, this involves counting context words. Say that my word of interest is *tea* and I have observed the following corpus:

(1a) He put down his cup of tea with force.

(1b) She drank a cup of tea.

Now I can compute a table of context word counts for *tea* from this corpus. First, I need to make some design decisions. For example, I consider the context of *tea* to be the sentence in which it occurs, and I count all lemmas except for stopwords. I then obtain the following table of counts:

cup	drink	force	put
2	1	1	1

I can reinterpret this table as a set of coordinates for the word *tea*, which put it at a point in a four-dimensional space whose dimensions are words, and the table of counts is now the vector of the word *tea*. This move allows us to formalize the pairwise meaning similarity of words, say, *tea* and *coffee*, as their proximity in space. This is usually defined as cosine similarity, the cosine of the angle of the vectors pointing to *tea* and *coffee* from the origin. For two vectors of numbers $\mathbf{a} = \langle a_1, \ldots, a_n \rangle$ and $\mathbf{b} = \langle b_1, \ldots, b_n \rangle$, the cosine is defined as

$$\cos(\mathbf{a}, \mathbf{b}) = \frac{\sum_i a_i b_i}{\sqrt{\sum_i a_i^2}\sqrt{\sum_i b_i^2}}.$$

Usually, a word is not represented through the original context word counts; rather, the vector is transformed to yield better similarity estimates. One such transformation is to automatically group context words (dimensions) whose import is similar in that they tend to appear in the same contexts.

Distributional models are not always probabilistic, but probabilistic versions exist. The most prominent of these is latent Dirichlet allocation (Blei et al. 2003). In its formulation for lexical semantics (Dinu & Lapata 2010), words are represented as probability distributions over semantic classes, which are probabilistic clusters of words that tend to co-occur in the same texts. The

probabilistic clusters in the probabilistic model play the same role as the dimensions in the non-probabilistic version described above.

Another variant of distributional models is based on the kind of supervised classifiers depicted in **Figure 2**. They are based on the insight that we can formulate contextual co-occurrence of words as a classification task: Given a word and a possible context word, do the two ever appear in the same context in a text? This is a binary classification task, where, for the pair ⟨*tea*, *cup*⟩, the answer is yes, according to the mini-corpus above. For the pair ⟨*tea*, *newt*⟩, the answer is no. The same corpus that can be used to count context words can also be used as training data to classify potential context words. The classifier's guess as to whether or not two words co-occur in a corpus is not particularly useful. Instead, we want the weights that the classifier learns (these are the weight vectors in **Figure 2*c***). The weights used to classify the potential context words of *tea* form a vector, which can be used as the distributional representation for *tea* in the same way as a count vector.

Distributional models are also called semantic spaces. The idea is that the coordinates of the word vectors are dimensions in some semantic space and, especially with regard to classifier-based vectors, word-embedding models.

4.3. Distributional Models and the Lexicon

Distributional models are widely used to study lexical meaning, as discussed in overview articles by Lenci (2008, 2018), Erk (2012), and Boleda (2020). Distributional models naturally give rise to a graded representation of word meaning because proximity in space comes in degrees. The dimensions of a distributional vector are not usually interpretable, so the standard way to interpret a word's vector is by retrieving its nearest neighbors in distributional space. These are the words that, according to the model, are most similar in meaning. The quality of a distributional model is usually evaluated against human experimental data, such as human ratings of word similarity (Finkelstein et al. 2001, Bruni et al. 2014, Hill et al. 2015, Gerz et al. 2016) and semantic relations (Baroni & Lenci 2011).

A main advantage of distributional models is their ability to compute representations for a large number of words automatically from corpus data. In fact, Baroni et al. (2014) consider the "vastness of lexical meaning" to be one of the main problems in semantics, the problem of obtaining and evaluating representations not only for a few chosen words but for the whole expanse of the lexicon.

Gehrke & McNally (2019) note that the distributional representation of a word is overspecified in the same way as the above-described feature representations of Zeevat et al. (2017) because the vector for a word is learned from all its uses, across all senses. For example, a vector for *orange* would contain traces of both orange-the-fruit and orange-the-color contexts.

4.4. Discussion

How should graded or probabilistic models for lexical semantics be evaluated? An approach with handcrafted features can make nuanced statements about a specific phenomenon, like the vague adjectives presented by Sutton (2015). But it is hard to scale such an approach up to the whole lexicon. To take the example of vague adjectives, making handcrafted representations for the whole lexicon would involve identifying all vague adjectives, categorizing them as involving one or more dimensions, and specifying the dimensions—a gigantic task. Distributional approaches naturally scale up and are evaluated against large test data sets. But if the model fails for a particular word, that alone usually does not tell us much about whether the underlying structure of a model is right or wrong. It looks like what is needed is some way to draw on the

advantages of both handcrafted features and automatically derived features. Indeed, there are several ways in which people are trying to build this bridge.

First, can distributional models be used to approximate the inferences that other frameworks of lexical meaning would make? Asher et al. (2016) explore whether it is possible to implement the type-theoretic formalism of Asher (2011) through a distributional model. And Zeevat (2019) notes that there are machine learning models that guess interpretable features from distributional vectors. This points to another possible way to automatically acquire the features used in handcrafted feature approaches: Objects have features, and they are mentioned in language, so there should be observable co-occurrences of grounded features and textual contexts. Erk (2016) uses this idea as the basis for her proposal for how a person might form an approximate notion of, say, *alligator*, even if they would not be able to pick an alligator out of a crowd.

Second, Baroni et al. (2010) and Baroni & Lenci (2010) have developed distributional models with interpretable features. Their idea was to define syntactic patterns that would extract interpretable phrases from the context of a word. Intriguingly, their models have much to say about affordances, ways in which people interact with objects—a book can be *read* or *published*, a motorcycle can be *ridden* or *parked*—but much less about the attributes of objects.

Third, the metaphor of meaning as a space, with meaning similarity as proximity in space, is applicable beyond corpus-derived spaces. It is also used, for example, in the conceptual spaces of Gärdenfors (2004), where the dimensions are interpretable. Herbelot & Copestake (2021) point out that co-occurrences of individuals and grounded properties can be learned in exactly the same way as co-occurrences of words in a distributional model. For this reason, the knowledge of a speaker about objects and their properties can be considered a space that yields a graded notion of similarity among properties, as well as similarity among individuals, in exactly the same way that similarities among words can be computed in a distributional model. Herbelot (2020) then completes the circle: She learns co-occurrences of individuals and grounded properties from a corpus of annotated pictures, but then transforms the space with a supervised classifier (as in the third class of distributional models described above), with very good results on word-similarity tasks.

5. GRADED AND PROBABILISTIC REPRESENTATIONS OF SENTENCE MEANING

If the representations of word meanings are graded or probabilistic and may be points in a semantic space, then what is the representation of a sentence? Recent proposals differ widely in their characterizations of sentence meaning, in their formalizations, and in the phenomena they cover. Still, two major strands of thought are emerging. One says that if words are points in semantic space, then sentences should be, too. The other represents an utterance as some kind of probabilistic elaboration of the situation that the utterance describes. A second, orthogonal way to classify (some) of the approaches below is by a focus on a lego view or a network view of sentence meaning. By "lego view," I mean a focus on defining lexical items with their combination potentials. By "network view," I mean an emphasis on the interconnectedness of all the words in a sentence and beyond, and mutual context influences.

5.1. Sentences in Semantics Space

Compositional distributional semantics represents words, phrases, and sentences as vectors (or higher-dimensional matrices, called tensors) in space. The representations are automatically acquired from corpus data in the same way as word vectors; they allow for compositional semantics construction in the same way as typed lambda calculus (Church 1940); and they are evaluated against human judgments of phrase and sentence similarity, in an analogous way to

word vectors. This strand of research is not explicitly probabilistic but rather falls under the umbrella of probabilistic-ish approaches that share the research goals discussed in Section 1.

Two papers (Clark et al. 2008, Baroni & Zamparelli 2010) started this strand of research. If words have representations that are vectors or tensors of the right shape, they can be combined compositionally using standard methods from linear algebra. To understand this idea in more detail, it is useful to look at the example of adjectives and nouns presented by Baroni & Zamparelli (2010). A standard way to describe adjectives in Montague semantics (Montague 1970, 1973) is as functions. If a noun, say, *apple*, denotes the set of all apples, then *red* is a function that, when applied to *apple*, maps the set of all apples to the smaller set of red apples. Now suppose the meaning of *apple* is instead a context vector, a characterization of the contexts in which *apple* appears. Then the meaning of *red* should be a function that maps vector to vector such that, when applied to the vector of *apple*, it yields a good approximation of the contexts in which *red apple* appears. Baroni & Zamparelli (2010) choose a simple linear function that can be characterized through a matrix of weights, and these weights can be learned using corpus data. Scaling this idea up, a verb with two arguments is represented as a cube. Combining it with one argument noun yields a matrix, and adding the second argument noun results in a vector.

Compositional distributional semantics, which defines vectors and tensors for all lexical items and defines combinability by the dimensionality of the tensors, takes the lego view of sentence meaning construction. The theoretical basis for this line of thought is discussed by Baroni et al. (2014). Grefenstette & Sadrzadeh (2011) develop the formalism for transitive verbs. Paperno et al. (2014) propose a simplified formulation for greater ease of learning. Clark et al. (2013) show that it is possible to handle complex structures, particularly relative clauses, in this framework. Muskens & Sadrzadeh (2018) draw a connection to dynamic semantics, and Sadrzadeh et al. (2018) integrate compositional distributional semantics into dynamic syntax.

Several studies have used the resulting phrase vectors to probe linguistic questions. Vecchi et al. (2011) study phrase vectors of semantically anomalous phrases, and Boleda et al. (2013) use phrase vectors to ask how nonintersective adjectives influence meaning as viewed through contexts: Does a *former bassist* occur in contexts that differ from those of a *male bassist*?

5.2. Probabilistic Elaborations of a Situation

The other main idea about how to integrate graded or probabilistic word meanings into sentence meaning is to say that an utterance describes a situation but does so only partially and probabilistically, such that the meaning of an utterance is associated with a probabilistic description of a world or situation. Cooper et al. (2015) treat situations as atomic entities that are categorized, or typed, by propositions, for example, *the type of situations where Kim is smiling*. They formalize this idea in a type-theoretic framework where situations have types that correspond to propositions. Conditional probabilities connect types: Given that *s* is a situation of the type *Kim is smiling*, how likely is it that *s* is also of the type *Kim is happy*? The authors build on a particularly rich type theory (Cooper 2021) in which some types are records and those records can include entries that are probabilities. Cooper et al. connect this framework to Larsson's (2013) perception classifiers (discussed in Section 4) and give a compositional semantics for a fragment of English. Sutton (2015) uses a very similar framework where the types of situations are derived from infons, information items from Situation Theory (Barwise & Perry 1999). He distinguishes described situations from discourse situations to connect someone's statement that *Mary is tall*, a discourse situation type, to probabilities of Mary's height in described situations. He focuses on vague adjectives, as mentioned above. The type-theoretic frameworks of both Cooper et al. and Sutton also take a lego view of sentence meaning constructions, with types defining the shapes of the lego bricks.

Another group of approaches uses probability distributions over worlds or situations (van Eijck & Lappin 2012, Goodman & Lassiter 2015, Emerson 2018, Erk & Herbelot 2021). Goodman & Lassiter (2015) interpret the literal listener of RSA (see Section 3) as an interaction of two probabilistic processes, a cognitive process that probabilistically generates worlds and a linguistic process that discriminates worlds where an utterance u is true from those where it is false. The probability prior of the world, $p(w)$, incorporates the cognitive process: Worlds are probabilistically generated by concepts that are Bayesian objects, a probabilistic language of thought (Goodman et al. 2015). The likelihood of an utterance given a world, $p(u|w)$, constitutes the linguistic judgment.

In order to scale this idea up, the generation of worlds, $p(w)$, needs to be more constrained by the utterance at hand. Emerson (2018) and Erk & Herbelot (2021) use probabilistic formulations that are overall very similar to those of Goodman & Lassiter (2015) but have a situation generation process that is driven by the words in the utterance: The situation is generated from material that is linked to the words. Both Emerson's (2018) and Erk & Herbelot's (2021) formulations have at their core a graphical model (as discussed in Section 2) that is a network of conceptual representations, where edges indicate interactions between word meanings. Thus, these approaches strongly foreground the network view of sentence meaning. Emerson uses a graphical model that corresponds to a logical form for a sentence (Copestake 2009) with a node for each predicate. The value of each node is a vector in distributional space. Neighboring vectors constrain one another through selectional preferences. Emerson (2020a) simplifies the probabilistic inference for this framework by using classifiers. The model is trained at a large scale on corpus data and evaluated on phrase-similarity data sets that are also used in compositional distributional semantics.

The graphical model presented by Erk & Herbelot (2021) has nodes that are concepts, again with edges indicating selectional constraints. Concepts are again associated with locations in semantic space. In this framework, the graphical model has additional nodes that are scenarios, knowledge of typical events and their participants, which constrain meaning in context even between words that are not connected through a semantic role. The concepts in the conceptual graph are connected to discourse referents in the utterance u, and probabilistically generate predicates describing the referents. Despite the similarity between this formalism and Emerson's, Erk & Herbelot (2021) use their model in a quite different way. They make use of handcrafted features and zoom in on individual sentences and their analysis (with simulations using probabilistic programming languages).

Chersoni et al. (2019) paint a very similar picture of interactions between word meanings and general event knowledge, so they also adopt the network view, but they do not use an explicitly probabilistic formalism. They associate predicates in a logical form with distributional vectors. A global graph of observed events and event participants represents knowledge of typical events. Chersoni et al. focus on incremental understanding, especially on how knowledge of typical events (the global event graph) influences the listener's expectations of which words will come next. The model is trained on corpus data and evaluated at large scale against phrase-similarity and selectional preference data sets.

Like Chersoni et al., McNally & Boleda (2017) (see also McNally 2017 and Gehrke & McNally 2019) associate predicates in a logical form with distributional vectors and do not use an explicitly probabilistic formalism. Their focus is on what it means to include such vectors in logical form. McNally and colleagues suggest that nouns denote kinds and that kinds have a descriptive meaning that can be approximated through vectors as "ersatz conceptual structures." Descriptive meaning is composed along with but separately from referential meaning.

Venhuizen et al. (2021), also like Chersoni et al. (2019), focus on incremental processing. They represent sentence prefixes as vectors whose dimensions are worlds. The weight on a world reflects the degree to which the sentence prefix could be completed in a way that would be true in the world.

The approach taken by Bernardy et al. (2018) (see also Bernardy et al. 2019a,b) is very different from the other approaches discussed above in that it is about the shape of lexical entries, rather than their content, and does not presume or learn any permanent lexicon. Given a small set of premises, Bernardy et al. probabilistically learn, for these premises alone, a semantic space, as well as a region representation of predicates in that semantic space, to match the premises. For example, if one premise is *All newts are dangerous*, then the region for *newt* must be included in the *dangerous* region. Bernardy et al. rely on probabilistic programming languages as the lingua franca for their formalization. Their approach asks how probabilistic representations in semantic space should be shaped to enable the right inferences: What should they look like for vague predicates? For quantifiers? For generics? Bernardy et al. (2019a) introduce a test bank for these and other phenomena.

5.3. Discussion

Lego view approaches can explain how the meaning of a sentence is composed out of the meanings of its components, and which combinations do and do not succeed. The network view can explore different influences on meaning in context (from selectional constraints, from event knowledge, and so on). Ideally, we want an approach that can do both, though how these viewpoints could be combined is unclear.

More generally, the approaches discussed in this section differ in the phenomena they place at the center. Compositional distributional semantics aims to model phrase meaning similarity. Compositionality is a main concern, as is corpus-based automated acquisition as a way to address the "vastness" of the lexicon (Baroni et al. 2014). Cooper and colleagues view semantic learning as the core problem, along with compositionality. Emerson, Chersoni and colleagues, and Erk and Herbelot want to explain word meaning in context and the complex interplay of words within an utterance. Bernardy and colleagues concentrate on the structure of semantic representations rather than on the content of the lexicon and propose structures for generics, quantifiers, and vague predicates. They are not the only ones to consider the interplay of quantifiers and lexical items: Emerson and Cooper and colleagues do, too. Again, ideally we want an approach that could address all of these phenomena and more, but not all existing formalizations are suited for addressing all phenomena.

This brings us, again, to the problem of scaling up, discussed in Section 4. It is important to scale up so as to be able to handle arbitrary lexical items, but it is also important to test for linguistic adequacy in depth, and many large-scale database evaluations give us only a blurry picture. So how can we have both? A possible solution is to first define a formalism that can be evaluated in depth on individual sentences and then implement it through corpus-derived vectors. Another possible solution is to follow the approach of Bernardy and colleagues: to abstract from lexical items, such that the structure of semantic representations, rather than their content, can be evaluated at a large scale.

All probabilistic approaches to sentence meaning have to engage with the question of what the nature of meaning is. Is meaning defined by textual context? Are lexical features "in the world" or "in the mind"? This problem is particularly complicated in sentence meaning representations, where a coherent notion of meaning is needed for lexical items and for the sentence as a whole. In compositional distributional semantics, the meaning-as-context view is a natural fit. In contrast,

the probabilistic situation approaches discussed above view meaning as being both "about the mind" and "about the world," but differ in the "nexus" (Pelletier 2017) between the two levels. For McNally (2017), nouns denote kinds, and that is the nexus. For Cooper et al. (2015), Goodman & Lassiter (2015), and Emerson (2018), the nexus is the act of categorizing or classifying, while for Erk & Herbelot (2021), it is the act of generating or imagining referents with particular properties.

Which phenomena should be described as semantic, and which as pragmatic? Is there a clear boundary? And what is the role of world knowledge? Goodman & Lassiter (2015) clearly separate conceptual and linguistic knowledge, and semantics and pragmatics, and Emerson (2018) makes the same separation, but in general it is not very clear. All distributional approaches pick up on whatever is in the context, which is everything that is reflected in speakers' utterances—semantics, pragmatics, world knowledge, and all. Also, Chersoni et al. (2019) and Erk & Herbelot (2021) explicitly include world knowledge in the shape of event knowledge.

6. CONCLUSION AND OUTLOOK

I have argued that graded and probabilistic approaches in semantics and pragmatics all share several core research problems:

- Gradience in the phenomena that they study. The lexicon is "fluid" (Kintsch 2007); pragmatics is "gooey" (Franke & Jäger 2016).
- The need to prove learnability. A listener can learn semantic representations and pragmatic reasoning from data.
- Empirical evaluation against experimental data or corpus data at scale.
- Scaling up to the full size of the lexicon.

The methods used are sometimes explicitly probabilistic and sometimes not, but as I have argued, the boundaries among probabilistic frameworks, classifiers in machine learning, and distributional approaches have been blurred because of new technical developments. These techniques have also provided new tools for semantics and pragmatics, which have been used both for easier probabilistic inference (e.g., Emerson 2020a) and to implement the idea that humans may learn shortcuts to complex probabilistic functions (e.g., White et al. 2020). Also, probabilistic programming language may grow into a possible lingua franca for describing probabilistic analyses (e.g., Goodman & Lassiter 2015, Bernardy et al. 2018).

Among graded and probabilistic approaches to lexical meaning and sentence meaning there are currently a multitude of voices, of directions, of formalizations. I have sketched the main questions that I think the field needs to figure out. The first is how to build models that can undergo two important modes of evaluation: (*a*) the so-called magnifying-glass evaluation of individual sentences, which lets us see exactly what we got right and what we got wrong about a phenomenon but which does not scale up, and (*b*) large-scale empirical evaluations that challenge a model with the whole breadth of language. The problem is that models that can perform large-scale evaluations are often not built in a way that allow for magnifying-glass evaluations—but we need them both. The other main question is whether there is any framework that can incorporate the network view of word meaning in context, all the interconnections of context influences, while still providing a lego view of lexical items that compose into a sentence representation.

These questions can be summarized as follows: To what extent is it possible to integrate these approaches that cover different phenomena? This question applies to Bayesian pragmatics as well: Is it possible to incorporate the mutual pragmatic reasoning over belief states with an account of sentence meaning that incorporates the full vastness of the lexicon? Having such an all-encompassing formalism would give us a better basis, both empirical and theoretical, to

revisit the old question of the semantics–pragmatics boundary. Vagueness has been described in probabilistic terms as involving either pragmatic reasoning over the interlocutor (Lassiter & Goodman 2017) or a stored lexical transformation (Sutton 2015)—which is the more parsimonious assumption? Lassiter & Goodman (2017) argue that a Bayesian pragmatic account avoids having to make assumptions about specialized semantic representations. But if adjectives always induce a transformation of the semantic space location of the noun, as Baroni & Zamparelli (2010) propose, vague adjectives might not need any special machinery. Generic expressions have also been analyzed both from the pragmatic reasoning point of view (Tessler & Goodman 2019) and from the point of view of the shapes of their semantic representations (Bernardy et al. 2019a, Emerson 2020b). Do we need both halves of the story—pragmatic reasoning to know what to represent, as well as a semantic formalism to know how the result of pragmatic reasoning fits in an overall sentence representation? I think that such a push toward integration of frameworks and extension of covered phenomena will give us many complex puzzles to solve.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

I thank the anonymous reviewer for their excellent suggestions. And many thanks for all the helpful discussions go to the participants in my fall 2021 seminar: Samuel Cantor, Yejin Cho, Gabriella Chronis, Venkat Govindarajan, Ellen Jones, Gabriela O'Connor, and Juan Diego Rodriguez.

LITERATURE CITED

Asher N. 2011. *Lexical Meaning in Context: A Web of Words*. Cambridge, UK: Cambridge Univ. Press

Asher N, Van de Cruys T, Abrusán M. 2016. Integrating type theory and distributional semantics: a case study on adjective–noun compositions. *Comput. Linguist.* 42(4):703–25

Baroni M, Bernardi R, Zamparelli R. 2014. Frege in space: a program for compositional distributional semantics. *Linguist. Issues Lang. Technol.* 9(6):5–110

Baroni M, Lenci A. 2010. Distributional memory: a general framework for corpus-based semantics. *Comput. Linguist.* 36(4):673–721

Baroni M, Lenci A. 2011. How we BLESSed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics*, pp. 1–10. Washington, DC: Assoc. Comput. Linguist.

Baroni M, Murphy B, Barbu E, Poesio M. 2010. Strudel: a corpus-based semantic model based on properties and types. *Cogn. Sci.* 34(2):222–54

Baroni M, Zamparelli R. 2010. Nouns are vectors, adjectives are matrices: representing adjective–noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (*EMNLP 2010*), pp. 1183–93. Washington, DC: Assoc. Comput. Linguist.

Barwise J, Perry J. 1999. *Situations and Attitudes*. Stanford, CA: Cent. Study Lang. Inf.

Benz A, Stevens J. 2018. Game-theoretic approaches to pragmatics. *Annu. Rev. Linguist.* 4:173–91

Bernardy JP, Blanck R, Chatzikyriakidis S, Lappin S. 2018. A compositional Bayesian semantics for natural language. In *Proceedings of the 1st International Workshop on Language Cognition and Computational Models*, pp. 1–10. Washington, DC: Assoc. Comput. Linguist.

Bernardy JP, Blanck R, Chatzikyriakidis S, Lappin S, Maskharashvili A. 2019a. Bayesian inference semantics: a modelling system and a test suite. In *Proceedings of the 8th Joint Conference on Lexical and Computational Semantics* (**SEM 2019*), pp. 263–72. Washington, DC: Assoc. Comput. Linguist.

Bernardy JP, Blanck R, Chatzikyriakidis S, Lappin S, Maskharashvili A. 2019b. Predicates as boxes in Bayesian semantics for natural language. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pp. 333–37. Turku, Finl.: Linköping Univ. Electron. Press

Blei DM, Ng A, Jordan MI. 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022

Boleda G. 2020. Distributional semantics and linguistic theory. *Annu. Rev. Linguist.* 6:213–34

Boleda G, Baroni M, Pham TN, McNally L. 2013. Intensionality was only alleged: on adjective–noun composition in distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics* (*IWCS 2013*): *Long Papers*, pp. 35–46. Washington, DC: Assoc. Comput. Linguist.

Bruni E, Tran NK, Baroni M. 2014. Multimodal distributional semantics. *J. Artif. Intell. Res.* 49:1–47

Chersoni E, Santus E, Pannitto L, Lenci A, Blache P, Huang CR. 2019. A structured distributional model of sentence meaning and processing. *Nat. Lang. Eng.* 25(4):483–502

Church A. 1940. A formulation of the simple theory of types. *J. Symb. Log.* 5(5):56–68

Clark S, Coecke B, Sadrzadeh M. 2008. A compositional distributional model of meaning. In *Proceedings of the 2nd Quantum Interaction Symposium* (*QI'08*), pp. 133–40. New York: ACM

Clark S, Coecke B, Sadrzadeh M. 2013. The Frobenius anatomy of relative pronouns. In *Proceedings of the 13th Meeting on the Mathematics of Language* (*MoL 13*), pp. 41–51. Washington, DC: Assoc. Comput. Linguist.

Cohen A. 1999. Generics, frequency adverbs and probability. *Linguist. Philos.* 22:221–53

Cohen A. 2003. Probabilistic approaches to semantics. In *Probabilistic Linguistics*, ed. R Bod, J Hay, S Jannedy, pp. 343–79. Cambridge, MA: MIT Press/Bradford

Cooper R. 2021. *From perception to communication: an analysis of meaning and action using a theory of types with records* (*TTR*). Unpubl. Ms. **https://github.com/robincooper/ttl/blob/master/ttl.pdf**

Cooper R, Dobnik S, Lappin S, Larsson S. 2015. Probabilistic type theory and natural language semantics. *Linguist. Issues Lang. Technol.* 10:1–43

Copestake A. 2009. Slacker semantics: why superficiality, dependency and avoidance of commitment can be the right way to go. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics* (*EACL 2009*), pp. 1–9. Washington, DC: Assoc. Comput. Linguist.

Degen J, Hawkins R, Graf C, Kreiss E, Goodman N. 2020. When redundancy is useful: A Bayesian approach to "overinformative" referring expressions. *Psychol. Rev.* 127(4):591–621

Dinu G, Lapata M. 2010. Measuring distributional similarity in context. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (*EMNLP 2010*), pp. 1162–72. Washington, DC: Assoc. Comput. Linguist.

Emerson G. 2018. *Functional distributional semantics: learning linguistically informed representations from a precisely annotated corpus*. PhD Thesis, Univ. Cambridge, Cambridge, UK

Emerson G. 2020a. Autoencoding pixies: amortised variational inference with graph convolutions for functional distributional semantics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3982–95. Washington, DC: Assoc. Comput. Linguist.

Emerson G. 2020b. Linguists who use probabilistic models love them: quantification in functional distributional semantics. In *Proceedings of the 2020 Probability and Meaning Conference* (*PoM 2020*), pp. 41–52. Washington, DC: Assoc. Comput. Linguist.

Erk K. 2012. Vector space models of word meaning and phrase meaning: a survey. *Lang. Linguist. Compass* 6(10):635–53

Erk K. 2016. What do you know about an alligator when you know the company it keeps? *Semant. Pragmat.* 9(17):1–63

Erk K, Herbelot A. 2021. How to marry a star: probabilistic constraints for meaning in context. arXiv:2009.07936 [cs]

Falkum IL, Vicente A. 2015. Polysemy: current perspectives and approaches. *Lingua* 157:1–16

Fernández R, Larsson S. 2014. Vagueness and learning: a type-theoretic approach. In *Proceedings of the 3rd Joint Conference on Lexical and Computational Semantics* (**SEM 2014*), pp. 151–59. Washington, DC: Assoc. Comput. Linguist.

Finkelstein L, Gabrilovich E, Matias Y, Rivlin E, Solan Z, et al. 2001. Placing search in context: the concept revisited. In *Proceedings of the 10th World Wide Web Conference*, pp. 406–14. New York: ACM

Frank MC, Goodman ND. 2012. Predicting pragmatic reasoning in language games. *Science* 336(6084):998–98

Franke M, Jäger G. 2016. Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Z. Sprachwiss.* 35(1):3–44

Gärdenfors P. 2004. *Conceptual Spaces*. Cambridge, MA: MIT Press

Gehrke B, McNally L. 2019. Idioms and the syntax/semantics interface of descriptive content versus reference. *Linguistics* 57(4):769–814

Gerz D, Vulic 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* (*EMNLP 2016*), pp. 2173–82. Washington, DC: Assoc. Comput. Linguist.

Goodman ND, Frank MC. 2016. Pragmatic language interpretation as probabilistic inference. *Trends Cogn. Sci.* 20(11):818–29

Goodman ND, Lassiter D. 2015. Probabilistic semantics and pragmatics: Uncertainty in language and thought. In *The Handbook of Contemporary Semantic Theory*, ed. S Lappin, C Fox, pp. 655–86. New York: Wiley-Blackwell. 2nd ed.

Goodman ND, Mansinghka VK, Roy D, Bonawitz K, Tenenbaum JB. 2008. Church: a language for generative models. In *In Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence* (*UAI 2008*), pp. 220–29. Arlington, VA: AUAI

Goodman ND, Stuhlmüller A. 2014. *The Design and Implementation of Probabilistic Programming Languages*. **http://dippl.org**

Goodman ND, Tenenbaum JB, Gerstenberg T. 2015. Concepts in a probabilistic language of thought. In *The Conceptual Mind: New Directions in the Study of Concepts*, ed. E Margolis, S Laurence, pp. 623–53. Cambridge, MA: MIT Press

Grefenstette E, Sadrzadeh M. 2011. Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (*EMNLP 2011*), pp. 1394–404. Washington, DC: Assoc. Comput. Linguist.

Grice P. 1975. Logic and conversation. In *Syntax and Semantics*, Vol. 3: *Speech Acts*, ed. P Cole, J Morgan, pp. 41–58. New York: Academic

Hacking I. 2001. *An Introduction to Probability and Inductive Logic*. Cambridge, UK: Cambridge Univ. Press

Herbelot A. 2020. Re-solve it: simulating the acquisition of core semantic competences from small data. In *Proceedings of the 24th Conference on Computational Natural Language Learning* (*CoNLL 2020*), pp. 344–54. Washington, DC: Assoc. Comput. Linguist.

Herbelot A, Copestake A. 2021. Ideal words: a vector-based formalisation of semantic competence. *Künst. Intell.* 35:271–90

Hill F, Reichart R, Korhonen A. 2015. SimLex-999: evaluating semantic models with (genuine) similarity estimation. *Comput. Linguist.* 41(4):665–95

Kamp H. 1975. Two theories about adjectives. In *Formal Semantics of Natural Language*, ed. E Keenan, pp. 123–55. Cambridge, UK: Cambridge Univ. Press

Kintsch W. 2007. Meaning in context. In *Handbook of Latent Semantic Analysis*, ed. T Landauer, D McNamara, S Dennis, W Kintsch, pp. 89–105. Mahwah, NJ: Erlbaum

Koller D, Friedman N. 2009. *Probabilistic Graphical Models: Principles and Techniques*. Adapt. Comput. Mach. Learn. Ser. Cambridge, MA: MIT Press

Larsson S. 2013. Formal semantics for perceptual classification. *J. Log. Comput.* 25(2):335–69

Lassiter D, Goodman N. 2017. Adjectival vagueness in a Bayesian model of interpretation. *Synthese* 194(10):3801–36

Lenci A. 2008. Distributional semantics in linguistic and cognitive research. *Riv. Linguist.* 20(1):1–31

Lenci A. 2018. Distributional models of word meaning. *Annu. Rev. Linguist.* 4:151–71

Liang P, Potts C. 2015. Bringing machine learning and compositional semantics together. *Annu. Rev. Linguist.* 1:355–76

Mcclelland J. 2013. Integrating probabilistic models of perception and interactive neural networks: a historical and tutorial review. *Front. Psychol.* 4:503

McCready E, Ogata N. 2007. Evidentiality, modality and probability. *Linguist. Philos.* 30:147–206

McMahan B, Stone M. 2015. A Bayesian model of grounded color semantics. *Trans. Assoc. Comput. Linguist.* 3:103–16

McNally L. 2017. Kinds, descriptions of kinds, concepts, and distributions. In *Bridging Formal and Conceptual Semantics: Selected Papers of BRIDGE-14*, ed. K Balogh, W Petersen, pp. 39–61. Düsseldorf, Ger.: Düsseldorf Univ. Press

McNally L, Boleda G. 2017. Conceptual versus referential affordance in concept composition. In *Language, Cognition, and Mind*, Vol. 3: *Compositionality and Concepts in Linguistics and Psychology*, ed. J Hampton, Y Winter, pp. 245–68. Berlin: Springer

Merin A. 1999. Information, relevance, and social decisionmaking: some principles and results of decision-theoretic semantics. In *Logic, Language, and Computation*, Vol. 2, ed. LS Moss, J Ginzburg, M de Rijke, pp. 179–221. Stanford, CA: Cent. Study Lang. Inf.

Monroe W, Hawkins RX, Goodman ND, Potts C. 2017. Colors in context: a pragmatic neural model for grounded language understanding. *Trans. Assoc. Comput. Linguist.* 5:325–38

Montague R. 1970. English as a formal language. In *Linguaggi nella societá e nella tecnica*, ed. B Visentini, pp. 189–224. Milan: Ed. Comunitá

Montague R. 1973. The proper treatment of quantification in english. In *Approaches to Natural Language*, ed. K Hintikka, pp. 221–42. Dordrecht, Neth.: Reidel

Muskens R, Sadrzadeh M. 2018. Static and dynamic vector semantics for lambda calculus models of natural language. *J. Lang. Model.* 6(2):319–51

Noble B, Breitholz E, Cooper R. 2020. Personae under uncertainty: the case of topoi. In *Proceedings of the 2020 Probability and Meaning Conference* (*PoM 2020*), pp. 8–16. Washington, DC: Assoc. Comput. Linguist.

Oaksford M, Chater N. 2007. *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*. Oxford, UK: Oxford Univ. Press

Paperno D, Pham NT, Baroni M. 2014. A practical and linguistically-motivated approach to compositional distributional semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Vol. 1: *Long Papers*, pp. 90–99. Washington, DC: Assoc. Comput. Linguist.

Parikh P. 2000. Communication, meaning, and interpretation. *Linguist. Philos.* 23(2):185–212

Pearl J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA: Morgan Kaufmann

Pelletier FJ. 2017. Compositionality and concepts—a perspective from formal semantics and philosophy of language. In *Language, Cognition, and Mind*, Vol. 3: *Compositionality and Concepts in Linguistics and Psychology*, ed. J Hampton, Y Winter, pp. 31–94. Berlin: Springer

Qing C, Franke M. 2015. Variations on a Bayesian theme: comparing Bayesian models of referential reasoning. In *Bayesian Natural Language Semantics and Pragmatics*, ed. H Zeevat, HC Schmitz, pp. 201–20. Cham, Switz.: Springer

Sadrzadeh M, Purver M, Hough J, Kempson R. 2018. Exploring semantic incrementality with dynamic syntax and vector space semantics. In *Proceedings of the 22nd Workshop on the Semantics and Pragmatics of Dialogue* (*SemDial 2018*), pp. 122–31. Aix-en-Provence, Fr.: AixDial. **http://semdial.org/anthology/Z18-Sadrzadeh_semdial_0016.pdf**

Schuster A, Stroessner C, Sutton P, Zeevat H. 2020. Stochastic frames. In *Proceedings of the 2020 Probability and Meaning Conference* (*PoM 2020*), pp. 78–85. Washington, DC: Assoc. Comput. Linguist.

Solt S. 2015. Vagueness and imprecision: empirical foundations. *Annu. Rev. Linguist.* 1:107–27

Sutton P. 2015. Towards a probabilistic semantics for vague adjectives. In *Bayesian Natural Language Semantics and Pragmatics*, ed. H Zeevat, HC Schmitz, pp. 221–46. Cham, Switz.: Springer

Tessler MH, Goodman ND. 2019. The language of generalization. *Psychol. Rev.* 126(3):395–436

Tuggy D. 1993. Ambiguity, polysemy and vagueness. *Cogn. Linguist.* 4(2):273–90

van Eijck J, Lappin S. 2012. Probabilistic semantics for natural language. In *Logic and Interactive Rationality Yearbook*, Vol. 2, ed. Z Christoff, P Galeazzi, N Gierasimczuk, A Marcoci, S Smets, pp. 17–35. Amsterdam: Amsterdam Dyn. Group

Vecchi EM, Baroni M, Zamparelli R. 2011. (Linear) maps of the impossible: capturing semantic anomalies in distributional space. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, pp. 1–9. Washington, DC: Assoc. Comput. Linguist.

Venhuizen NJ, Hendriks P, Crocker MW, Brouwer H. 2021. Distributional formal semantics. arXiv:2103.01713 [cs]

White J, Mu J, Goodman N. 2020. Learning to refer informatively by amortizing pragmatic reasoning. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, pp. 994–1000. Austin, TX: Cogn. Sci. Soc.

Zeevat H. 2015. Perspectives on Bayesian natural language semantics and pragmatics. In *Bayesian Natural Language Semantics and Pragmatics*, ed. H Zeevat, HC Schmitz, pp. 1–24. Cham, Switz.: Springer

Zeevat H. 2019. From semantic memory to semantic content. In *Language, Logic, and Computation*, ed. A Silva, S Staton, P Sutton, C Umbach, pp. 312–29. Berlin: Springer

Zeevat H, Grimm S, Hogeweg L, Lestrade S, Smith E. 2017. Representing the lexicon. In *Bridging Formal and Conceptual Semantics: Selected Papers of BRIDGE-14*, ed. K Balogh, W Petersen, pp. 153–86. Düsseldorf, Ger.: Düsseldorf Univ. Press