Functional Distributional Semantics: Learning Linguistically Informed Representations from a Precisely Annotated Corpus

Guy Emerson

The aim of distributional semantics is to design computational techniques that can automatically learn the meanings of words from a body of text. As McNally (2017) and Lenci (2008, 2018) have argued, distributional representations can be used as surrogates for conceptual representations – but crucially, they can be calculated concretely. Used in this way, distributional data allows us to develop and test linguistic theories. The twin challenges are: how do we represent meaning, and how do we learn these representations? The current state of the art is to represent meanings as vectors – but vectors do not correspond to any traditional notion of meaning. In particular, there is no way to talk about *truth*, a crucial concept in logic and formal semantics.

In this thesis, I develop a framework for distributional semantics which answers this challenge. The meaning of a word is not represented as a vector, but as a *function*, mapping entities (objects in the world) to probabilities of truth (the probability that the word is true of the entity). Such a function can be interpreted both in the machine learning sense of a classifier, and in the formal semantic sense of a truth-conditional function. This simultaneously allows both the use of machine learning techniques to exploit large datasets, and also the use of formal semantic techniques to manipulate the learnt representations. I define a probabilistic graphical model, which incorporates a probabilistic generalisation of model theory (allowing a strong connection with formal semantics), and which generates semantic dependency graphs (allowing it to be trained on a corpus).

Empirically, I demonstrate the feasibility of this approach by training a model on WikiWoods, a parsed version of the English Wikipedia, and evaluating it on three tasks. The results indicate that the model can learn information not captured by vector space models. Theoretically, I demonstrate the expressiveness of this approach, by showing how it can naturally model semantic composition, context dependence, and generalised quantifiers, with Bayesian inference playing a crucial role. The ability to model these phenomena is a clear advantage over other approaches to distributional semantics.

Chapter 1: Between Linguistics and Machine Learning

If we are to take lexical semantics seriously – that is, to have a theory that can model the meanings of words, including all their subtle connotations – then we cannot hope to write down all the details by hand. Traditional techniques are time-consuming, and variations of meaning difficult to pin down. Data-driven techniques are necessary to move from an abstract semantic theory to a fleshed-out model of a real language.

With this perspective in mind, in this chapter I situate my thesis against a background of work on distributional semantics and model-theoretic semantics. For distributional semantics, I discuss the early motivations (for example: Harris, 1954; Firth, 1951, 1957; Spärck-Jones, 1964), modern linguistic interest (for example: McNally, 2017; Lenci, 2008, 2018), and modern vector space models (for example: Turney and Pantel, 2010; Mikolov et al., 2013). For model-theoretic semantics, I give a basic introduction (following expositions such as: Cann, 1993; Allan, 2001; Kamp and Reyle, 2013), and cover three important developments relevant to the thesis: neo-Davidsonian event semantics (Davidson, 1967; Parsons, 1990), situation semantics (Barwise and Perry, 1983), and Dependency Minimal Recursion Semantics (DMRS) (Copestake et al., 2005; Copestake, 2009).

As well as summarising existing work, this chapter also serves to emphasise two core ideas which I build on in this thesis. One idea is that the meaning of a word should be represented by a *truth-conditional function*, a mapping from entities (objects in the world) to truth values (either true or false). The second core idea is that meanings of sentences can be represented as *semantic dependency graphs* – in particular, DMRS graphs. Compared to other sentence representations used in formal semantics, dependency graphs are more convenient for machine learning models – in particular, for *probabilistic graphical models*, which also use graph structures.

Chapter 2: Modelling Meaning in Distributional Semantics and Model-Theoretic Semantics

In this chapter, I discuss the goals of a semantic theory. I take the ultimate aim to be to characterise the meanings of all utterances in a language. To make this aim more precise, I elaborate on several aspects of meaning which could be considered crucial. For each aspect, I identify a plausible goal for a semantic theory, lay out out the space of possible theories, place existing work in this space, and evaluate which approaches seem most promising. This chapter motivates the framework developed in the rest of the thesis.

One goal for a semantic theory is to explain how language relates to the world, – in Harnad (1990)'s terms, to *ground* the semantics. Pure distributional semantics is not grounded, since it only uses textual data. I review three possible approaches to grounding a distributional model: concatenating distributional vectors with grounded vectors (for example: Bruni et al., 2011), mapping distributional representations to grounded representations (for example: Lazaridou et al., 2014), and jointly learning semantic representations from both distributional and grounded data (for example: Feng and Lapata, 2010). I argue that joint learning is the right approach, because such models are connected to grounded data from the outset, rather than trying to make such connections after the fact.

Model-theoretic semantics can be grounded, if individuals in a model structure are real-world individuals. However, one goal for a semantic theory is to be able to generalise to new situations, where the individuals have not previously been observed. To achieve this, we need to make a clear distinction between a *concept* (the meaning of a word) and a *referent* (an individual in an extension), as noted by Ogden and Richards (1923) and others. I discuss formal solutions to this problem, including representing a concept as a set of properties (for example: Arnauld and Nicole, 1662; Jones, 1911) or as a function from possible worlds to extensions (Carnap, 1947), but point out philosophical objections (Wittgenstein, 1953) and psychological objections (Rosch, 1975, 1978). I conclude that we need rich lexical representations, which machine learning models might provide. However, I argue that vector space models do not have enough structure to relate concepts to referents. I identify two promising alternatives: representing a concept by a region of space (for example: Gärdenfors, 2000, 2014), or by a binary classifier (for example: Larsson, 2013), both of which have a clear notion of truth.

I then discuss more specific goals for lexical semantics. One goal is to account for *vagueness*: where individuals can fall along a spectrum (such as size or colour), many predicates do not have a strict cutoff between truth and falsehood. I discuss two existing solutions for model theory: fuzzy truth values (for example: Zadeh, 1965, 1975) and probabilistic uncertainty about truth values (for example: Lassiter, 2011; Sutton, 2015). I also discuss probabilistic vector space models (for example: Vilnis and McCallum, 2015), which at first sight seem to deal with vagueness. However, I argue that they fail to do so, because they do not distinguish concepts and referents, and so have no notion of truth.

A second challenge for lexical semantics is *polysemy*. A particular difficulty is that there are often many ways we can divide usages of a word into a discrete set of senses (as discussed by: Kilgarriff, 1997, 2007; Hanks, 2000; Erk, 2010). Therefore, one goal for a semantic theory is to deal with polysemy without finite sense inventories. Indeed, Ruhl (1989) argues that even highly frequent terms with many apparent senses, such as *bear* and *hit*, can be assigned a single underspecified meaning. I argue that representing meaning by a region of space or a classifier is compatible with this view, but this approach has been little explored in both model theory and distributional semantics. The challenge would be to accurately represent such meanings without overgeneralising to cases where they wouldn't be used.

A final goal for lexical semantics is to model relations between words, such as *hyponymy*. Model theory captures this easily, when one extension is a subset of another. Distributional semantics faces more of a challenge. I discuss several approaches, but I note that it has been empirically shown that hyponymy is difficult to detect in vector space models (Bordea et al., 2016; Camacho-Collados et al., 2018) – state-of-the-art systems augment vector space models by searching for string patterns (e.g. *X such as Y*), following Hearst (1992). I argue that, to model hyponymy, we must move away from vector space models. As discussed by Erk (2009a,b) and Gärdenfors (2014, §6.4), modelling meaning as a region of space provides a natural definition of hyponymy.

I then move on to discuss sentence-level semantics, beginning with *compositionality*. The goal is to be able to derive the meaning of a sentence from its parts, so that we can generalise to new sentences. I note that compositionality is a strength of model theory (including for MRS, see: Copestake et al., 2001; Copestake, 2007), and I present a theoretical argument that vector space models produce lossy compressions of sentence meaning, rather than full representations.

However, arguing against vector representations does not tell us how we *should* represent sentence meanings. I therefore turn to logic, which formalises how sentences can be used to build chains of reasoning. One goal for semantics is to support the notions of *truth* and *entailment*. Again, I note that this is a strength of model theory. In contrast, while there has been work on using distributional vectors in a logical system (for example: Garrette et al., 2011; Beltagy et al., 2016), as well as developing neural network models with logical structure (for example: Andreas et al., 2016a,b), there has not been work on directly learning logical representations from distributional data. This thesis attempts to do that.

The flipside of compositionality is *context dependence*. Following Recanati (2012), I use *standing meaning* to refer to the context-independent meaning of an expression, and *occasion meaning* to refer to the context-dependent meaning of an expression in a particular occasion of use. One goal for semantics is to model how occasion meanings can be derived from standing meanings. This is challenging for both model theory and distributional semantics, but I identify probabilistic models as a promising approach, with initial steps taken in both formal semantics (for example: Lassiter and Goodman, 2015; Goodman and Frank, 2016) and distributional semantics (for example: Bražinskas et al., 2018).

The last goal I consider is for semantic representations to be *learnable* based on observed data. I note that there is little work on machine learning of model theory (although some limited datasets exist, such as: Young et al., 2014; Hürlimann and Bos, 2016). In contrast, distributional models are almost all learnable, but I highlight the tradeoff between expressiveness and learnability.

I conclude the chapter by reviewing existing approaches, in the light of the above goals. I consider extensions of vector space models, hybrid distributional-logical models (for example: Lewis and Steedman, 2013; Erk, 2016), the type-driven tensorial framework (Coecke et al., 2010; Baroni et al., 2014), and probabilistic semantic models (for example: Goodman and Lassiter, 2015; Cooper et al., 2015). For each, I highlight strengths and weaknesses, to motivate my own approach.

Chapter 3: Formal Framework of Functional Distributional Semantics

Having set the scene, in this chapter I introduce my framework. I begin by considering model theory, contrasting two formalisations of predicates: as extensions (sets of invididuals) or as truth-conditional functions (mapping from individuals to truth values). If individuals are atomic, converting between these two formalisations is trivial, I point out that, if inviduals are structured, the two may have rather different representations. To clarify the discussion, I introduce the term *pixie* to refer to the featural representation of an individual. Pixies can be considered as points in a *semantic space*.

I then propose a probabilistic generalisation of a model structure, allowing truth values to be uncertain, and allowing features of individuals to be uncertain. This generalisation lays the groundwork for the two core ideas introduced in Chapter 1. First, I define *semantic functions*, which are probabilistic truth-conditional functions: they map pixies to probabilities of truth. An example is given in Fig. 1, showing how the probabilities allow us to perform Bayesian inference.

I then examine semantic functions in more detail. I first note that representing lexical meaning as a semantic function places an emphasis on classification rather than generation, and I point to work in psychology supporting this view, with an extended discussion of Wong et al. (2018), who investigate knowledge of two forms of lowercase *g*: looptail ⟨g⟩, and opentail ⟨g⟩. They show that while English speakers can easily recognise both forms, they struggle to produce looptail ⟨g⟩.

Next, I compare semantic functions to regions of space, which are the two approaches to representing conceptual structure that I identified as promising in Chapter 2. I argue that they are two views of the same thing. For a non-probabilistic model, the correspondence is exact, but for a probabilistic model, the probabilities are in slightly different places. Knowing about a region of space gives us *global* information, while knowing about truth values for particular pixies is *local* information. To bridge this gap, we can parametrise a distribution over regions as a semantic function plus a covariance function (a mathematical technique reminiscent of Gaussian processes: Rasmussen and Williams, 2006). This maintains the coherence of truth values enforced by a region of space, and also allows efficient representations, because the covariance function could be shared across predicates.

Finally, I ask how the probabilities are supposed to be interpreted, building on previous work in philosophy of language (Lassiter, 2011; Sutton, 2017). While uncertainty about the world can be resolved via experiment, uncertainty about truth values does not have an obvious resolution, since the use of language is simply a convention (in the sense of: Millikan, 1998). However, no matter how precise a language convention might be, a speaker will always have to generalise to new situations. I argue that this need for generalisation leads to uncertainty.

Having introduced and discussed semantic functions, I turn to the second core idea, of using semantic dependency graphs. I define a probabilistic graphical model, following the probabilistic generalisation of a model structure introduced at the beginning of the chapter. Pixie-valued random variables represent individuals, and binary-valued random variables represent truth values. This graphical model defines a joint distribution over the pixie variables (intuitively, this distribution represents world knowledge), and conditional distributions over truth values given pixies (intuitively, these distributions represent lexical semantic knowledge).

Extending this graphical model to distributional semantics requires adding predicate-valued random variables, to represent what we observe in our corpus. This gives the graphical model in Fig. 3, which is the core of the framework of Functional Distributional Semantics. The basic assumption is that observed DMRS graphs are true of some situation. While we don't observe the individuals in each situation, we can leverage structural linguistic knowledge in the form of semantic dependency graphs, as illustrated in Fig. 2. After observing many such dependency graphs, we can jointly learn about what kinds of situations are likely to exist, as well as how these situations are likely to be described.

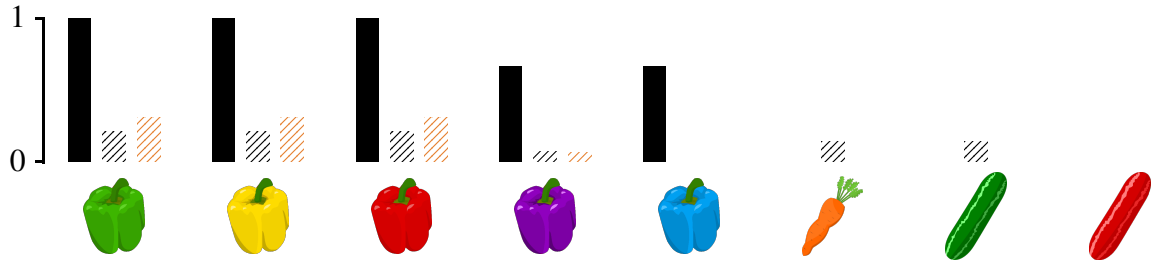

Figure 1: An example of Bayesian inference. We are interested in: an individual X, represented by some pixie $x \in \mathcal{X}$; and the truth value T of the *pepper* predicate for that individual. Each image above is a pixie.
Solid bars: the semantic function $\mathbb{P}(T = \top \mid X = x)$ represents a speaker's belief about whether each pixie x can be considered to be a pepper. This can be seen as a probabilistic truth-conditional function.
Black shaded bars: the prior $\mathbb{P}(X = x)$ represents the speaker's belief about an individual, based on their world knowledge. It encodes how much they expect to observe an individual with particular features.
Orange shaded bars: the posterior $\mathbb{P}(X = x \mid T = \top)$ represents their belief about an individual, if they know the *pepper* predicate is true of it. The probability mass is split between the pepper pixies, but skewed towards typical colours, and excluding colours believed impossible. This can be seen as a probabilistic extension.

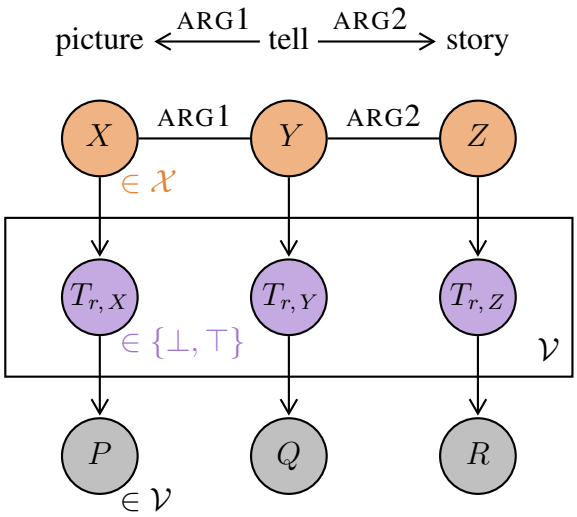

Figure 2: A simplified DMRS graph, which could be generated by Fig. 3 below. Such graphs are observed during training.

Figure 3: The probabilistic graphical model for Functional Distributional Semantics. Each node denotes a random variable (with possible values indicated in the first column). The *plate* (rectangle in the middle row) denotes repetition of nodes. Edges between nodes indicate probabilistic dependence.
Top row: individuals represented by pixie-valued random variables X, Y, Z, jointly distributed according to the DMRS links.
Middle row: for each individual, each predicate r in the vocabulary $\mathcal{V}$ is randomly true or false, with probability according to the predicate's semantic function.
Bottom row: for each individual, we randomly generate one predicate, out of all predicates true of the individual. This generates a DMRS graph as in Fig. 2. Only these nodes are observed.

I conclude the chapter by evaluating my framework against the goals given in Chapter 2. First, I note that it makes a clear distinction between concepts and referents. Although it is trained only on corpus data, I argue that it can be grounded in a more principled way than a vector space model, because the same semantic space could be used for both grounded data and latent pixies. Turning to lexical semantic goals, I first note that vagueness is built into the definition of a semantic function. Meanwhile, the model-theoretic definition of hyponymy in terms of subsets can be straightforwardly transferred. Polysemy is harder to formally define, since the goal is to avoid finite sense inventories. However, I argue that a semantic space allows a more fluid notion of sense, with Ruhl (1989)'s underspecified senses formalised as semantic functions taking high values over a large region of space. Turning to sentence-level goals, I first note that DMRS provides a mechanism for semantic composition. Meanwhile, truth values are built into the probabilistic graphical model, and the separation of predicates and pixies allows a natural separation of standing meaning and occasion meaning. These last two points will be discussed further in Chapter 4. Finally, for the goal of learnability, I note that the latent variables make learning more challenging, but I will present tractable learning algorithms in Chapter 5.

Chapter 4: From Bayesian Inference to Logical Inference

In this chapter, I explain how the logical structure of the framework is useful. I first show how Bayesian inference can provide an account of context-dependent meanings. This account maintains the intuition of Searle (1980) and Elman (2009), that language cannot be understood independently of world knowledge. However, this is done using a precisely defined mathematical mechanism: world knowledge is represented as a prior distribution over situations, standing meaning is represented using DMRS graphs and semantic functions, and occasion meaning is calculated as the posterior distribution over situations given the truth of a DMRS graph.

This account of context dependence can be easily extended to other kinds of context, because probabilistic graphical models provide a flexible way to extend a distribution to include more variables. I argue that, compared to classical approaches to context dependence (for example: Kaplan, 1979, 1989; Recanati, 2012), my approach is both more precise in its mathematical mechanism, and more general in its dependence on arbitrary context. Furthermore, I show how this approach to context dependence allows predicates to mutually disambiguate one another, because the occasion meanings of all predicates can be calculated jointly.

I then discuss how this kind of context dependence interacts with semantic composition. Given two DMRS graphs and appropriate prior distributions over situations, the occasion meaning of each graph is its posterior distribution over situations. I explain that composing the two graphs produces a new posterior distribution over larger situations, but this new posterior is not the same as independently combining the posteriors of the two graphs. This is because, as soon as the graphs are composed, the random variables are connected, and so the new posterior distribution depends on the entire composed graph. I relate this to what McNally and Boleda (2017) term *conceptually afforded* composition and *referentially afforded* composition. Composition of DMRS graphs relies on knowledge of concepts stored in the lexicon, and how they can combine. In contrast, composition of occasion meanings relies on knowledge of referents in a situation.

The above discussion has been about Bayesian inference over pixie-valued random variables. In the last part of the chapter, I turn to Bayesian inference over truth-valued random variables, showing how this can be used for logical inference. I set up a correspondence between logical propositions and statements about conditional probabilities. For example, a universally quantified proposition corresponds to a conditional probability of 1, while a existentially quantified proposition corresponds to a conditional probability greater than 0. Furthermore, I prove an equivalence with traditional syllogistic logic. Removing bound variables in a traditional logic corresponds to marginalising out random variables in my probabilistic approach. This correspondence will be generalised and further explored in Chapter 7.

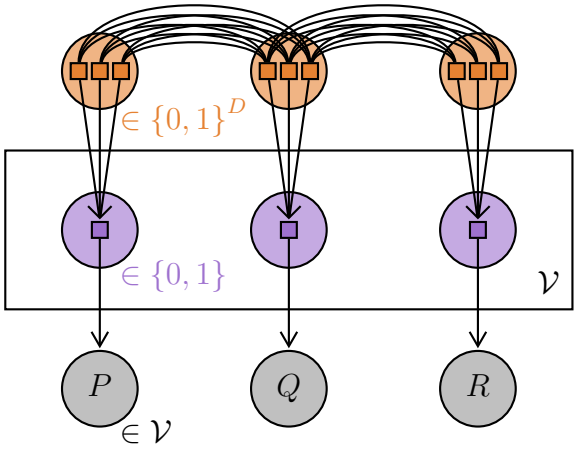

Figure 4: This neural network model implements the probabilistic graphical model in Fig. 3. Each square represents a binary unit (with D units per pixie).
Top row: pixies are binary-valued vectors, together forming a Restricted Boltzmann Machine. For each semantic dependency, there are connections between the pixies' units, determining how likely it is that pairs of units are active at the same time.
Middle row: each semantic function is a one-layer feedforward network, with a single output in the range $[0, 1]$, which is the probability that the predicate is true.
Bottom row: true predicates are weighted by frequency, and a single predicate is generated.

Chapter 5: Implementation and Inference Algorithms

The discussion in Chapters 3 and 4 has been relatively theoretical. A probabilistic graphical model only gives constraints on a distribution, rather defining a specific distribution. In this chapter, I explicitly construct a distribution that implements this graphical model, and I derive tractable inference algorithms.

The most crucial part of this chapter is the description of the model architecture, illustrated in Fig. 4. Because of the combination of directed and undirected edges in the graphical model (necessary for the equivalence with model theory), there are no existing neural network models that can be directly adapted for this purpose. Nonetheless, I have tried to keep the network architecture as simple as possible, so as not to introduce additional challenges.

I take pixies to be sparse binary-valued vectors. Intuitively, each dimension represents a different feature. Sparse representations have been shown to be beneficial in NLP, both for applications and for interpretability of features (for example: Murphy et al., 2012; Faruqui et al., 2015).

To define the distribution over situations, I use a Restricted Boltzmann Machine (RBM) (Smolensky, 1986; Hinton et al., 2006; Hinton, 2010), as adapted by Swersky et al. (2012) for sparse vectors. This is illustrated in the top row of Fig. 4. Each type of semantic dependency (such as ARG1 or ARG2) has a parameter matrix, which determines the strength of association between features of the linked pixies. To define the semantic functions, I use single-layer feedforward neural networks. This is illustrated by the connections between the top and middle rows of Fig. 4. Each semantic function has a parameter vector, which determines the strength of association with each feature. Finally, to generate predicates based on truth values (the bottom row of Fig. 4), I use a hard-coded method, for simplicity. The probability of generating a predicate is weighted according to the probability of truth and the frequency of the predicate.

After defining the model, I discuss the fact that the model only implements soft constraints on semantics, because probabilities of truth are never exactly 0 or 1. However, I point out that the high dimensionality of the space means that interesting subspaces are very small compared to the whole space, with the result that, in order to be useful, semantic functions must in practice be close to step functions. This makes them look more like classical truth-conditional functions taking only 0 and 1 as values.

The rest of the chapter is dedicated to how to train and use the model. The main challenge is the large number of latent variables: when training a model on text, we only observe predicates, not truth values or pixies. This means that, for every observed predicate, there is a latent pixie which the predicate is true of. Furthermore, for each latent pixie there is also a latent truth value for every other predicate in the vocabulary.

I propose training the model using gradient descent, a standard optimisation algorithm in machine learning. The aim is to maximise the probability of observing the training data, based on the parameters of the RBM and of the semantic functions. I give the gradients of the log-likelihood with respect to these parameters, along with both full derivations and intuitive explanations.

However, exactly calculating the gradients requires summing over the entire semantic space. For a high-dimensional space, this is infeasible. To make training tractable, I adapt two approximate inference techniques.

The first technique is Markov Chain Monte Carlo (MCMC). This approximates the intractable sum by sampling a small number of points, which should be representative of the whole distribution. This is often used for RBMs (Hinton, 2002, 2010). In our case, we need to sample from three distributions: pixies not conditioned on any observations, pixies conditioned on observed predicates, and truth values conditioned on latent pixies. For the first distribution, I propose using belief propagation (see: Yedidia et al., 2003), as applied to RBMs by Swersky et al. (2012). For the two conditional distributions, I propose using the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970).

These MCMC algorithms are guaranteed to provide unbiased estimates, but they may require many iterations, which means that they can be slow in practice. The second approximate inference technique I present is Variational Inference. The idea is to directly approximate the distribution we would like to calculate, and then optimise this approximation. I

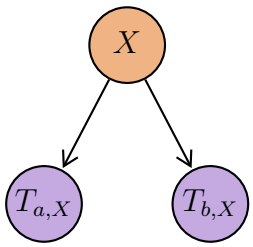

Figure 5: Logical inference for a single pixie X. We would like to know if the predicate b is true of X, given that the predicate a is true of X. This can be cast as the conditional probability $\mathbb{P}\left(T_{b,X} = \top \mid T_{a,X} = \top\right)$.

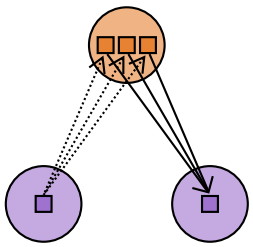

Figure 6: Variational inference for Fig. 5. The dotted lines indicate mean-field variational inference, and the solid lines indicate approximate inference by applying a semantic function to the mean-field vector.

use a *mean-field* approximation, where each dimension has an independent probability of being active, and all of these probabilities are jointly optimised. I derive an algorithm similar in spirit to Expectation Propagation (Minka, 2001), but I note that Expectation Propagation cannot be applied directly, because my model is not in the exponential family.

Finally, I explain how these approximate inference techniques can be used not just for training, but also for making tractable the calculations proposed in Chapter 4. In particular, I explain how the mean-field vectors produced by Variational Inference can be seen as approximate occasion meanings, and I explain how they can be used to perform approximate logical inference.

Chapter 6: Experiments

In this chapter, I bring together the work of the previous chapters, and test my framework empirically. For my training corpus, I use WikiWoods, a parsed version of the English Wikipedia, provid DMRS graphs for 55m sentences (Flickinger et al., 2010; Solberg, 2012). I describe my pre-processing steps, extracting simplified DMRS graphs using Pydmrs (Copestake et al., 2016), and lemmatising out-of-vocabulary items using Morphy (Fellbaum, 1998; Bird et al., 2009). I also transform DMRS graphs involving the copula, to enforce coreference between the subject and object.

I discuss all hyperparameters which must be set before training begins. These include decisions about the model architecture (such as dimensionality), decisions about the objective function (such as regularisation, to avoid overfitting) and decisions about optimisation algorithm (such as the learning rate).

I also discuss how the model parameters should be initialised, as this can reduce the training time. To initialise the semantic function parameters, I propose using Random Positive-only Projections, a random-indexing technique for producing count vectors (QasemiZadeh and Kallmeyer, 2016). To intialise the RBM parameters based on the semantic function parameters, I propose using mean-field vectors, treating each pixie as independent. This allows the RBM matrices to be intialised as sums of outer products of mean-field vectors.

After training a model, I turn to evaluation. I note that finding a good evaluation task is far from obvious. Simple similarity tasks do not require structured semantic representations like dependency graphs, while tasks like textual entailment require a level of coverage beyond the scope of this thesis. As well as evaluating on lexical similarity datasets, I also consider two additional datasets which allow us to explore approaches to semantic composition: one involves measuring similarity in context, and the other involves composition of relative clauses.

For lexical similarity, I evaluate on four datasets: SimLex-999 (Hill et al., 2015), SimVerb-3500 (Gerz et al., 2016), MEN (Bruni et al., 2014), and WordSim-353 (Finkelstein et al., 2001; Agirre et al., 2009). The range of datasets allows us to contrast *similarity* and *relatedness*. For example, *painter* and *painting* are related (since a painter paints paintings), but they are unlikely to be true of the same individuals, and the individuals they are true of are unlikely to share features. In contrast, *painter* and *artist* are similar. Vector space models tend to conflate these two notions. In contrast, my results demonstrate that my model captures similarity without conflating it with relatedness – it outperforms state-of-the-art vector space models on similarity datasets, while showing low correlation for relatedness.

To investigate context dependence, I used the dataset of Grefenstette and Sadrzadeh (2011), which consists of pairs of subject-verb-object triples, where only the verb varies in each pair – for example, (*table*, *show*, *result*) and (*table*, *express*, *result*). Each pair was annotated for similarity. The performance of my model matches the best model Grefenstette and Sadrzadeh consider. Furthermore, an ensemble model (combining my model and a vector space model) matches the improved model of Grefenstette et al. (2013), despite using less training data. It is also important to note that the fact that the ensemble outperforms both the semantic function model and the vector space model shows that the two models have learnt different kinds of information.

Finally, to investiage semantic composition, I used the RELPRON dataset (Rimell et al., 2016). This consists of a set of *terms*, each paired with up to ten *properties*. Each property is a short phrase, consisting of a hyperonym of the term, modified by a relative clause with a transitive verb. For example, a *telescope* is a *device that astronomers use*, and a *saw* is a *device that cuts wood*. The task is to identify the properties for each term.

A model that uses relatedness can perform fairly well on this dataset – for example, *astronomer* can predict *telescope*, without knowing what relation there is between them. However, the dataset also includes lexical *confounders* – for example, a *document that has a balance* is a financial *account*, not the quality of *balance* (not falling over). The lexical overlap means that the confounders confuse vector space models, including the models that Rimell et al. tested. Overall, my model performs worse than vector addition, perhaps as expected, since it does not capture relatedness. However, an ensemble (combining my model and a vector space model) performs better than either model alone. In particular, the ensemble performs much better than the vector space model on the lexical confounders. In fact, the ensemble weights indicate that the semantic function model has effectively taken over responsibility for deciding if the head noun is a hyperonym of the term, while the vector space model can better detect relatedness between the other noun and the term. To my knowledge, this is the first system that manages to improve both overall performance as well as performance on the confounders.

Chapter 7: Quantifiers and First-Order Logic

In this chapter, I take the logical equivalence presented in Chapter 4, and extend it to deal with multiple quantifiers, thereby allowing us to handle arbitrary propositions in first-order logic. This is an important strength compared to current distributional semantic models.

I first provide a background on generalised quantifiers (Barwise and Cooper, 1981; Van Benthem, 1984), and scope trees as illustrated in Fig. 7. This gives the classical non-probabilistic view which I aim to generalise probabilistically. A generalised quantifier assigns truth based on the cardinality of the restriction and the cardinality of the intersection of the restriction and body. I define a probabilistic quantifier to assign a probability of truth, based on the conditional probability of the body given the restriction – this conditional probability is exactly the ratio of the two classical cardinalities. Finally, just as a logical proposition can combine several quantifiers in a scope tree, we can define multiple random variables to create a probabilistic scope tree, as illustrated in Fig. 8 (compared to Fig. 7, α is *picture*, β is *tell*, γ is *story*, and the scope tree is upside-down). Quantifying over a logical variable corresponds to marginalising out a random variable.

I then show how a probabilistic approach provides a natural account of vague quantifiers (such as *few* and *many*). This is in the spirit of classical accounts (for example: Partee, 1988), but made more precise by using probability to model an underspecified threshold. Furthermore, this account extends to *generic* quantification (for example, *dogs bark* and *ducks lay eggs*). Generics are ubiquitous in natural language, but they are challenging for classical models, because the truth conditions seem to depend heavily on context (for example: Carlson, 1977; Carlson and Pelletier, 1995; Leslie, 2008; Herbelot, 2010). I show how my approach is compatible with Tessler and Goodman (2016)'s account of generics using Rational Speech Acts (RSA), a Bayesian approach to pragmatics (Frank and Goodman, 2012; Goodman and Frank, 2016).

In the last part of the chapter, I explain how precise quantifiers can be meaningfully combined with vague predicates, and I relate this chapter to the simpler account given in Chapter 4.

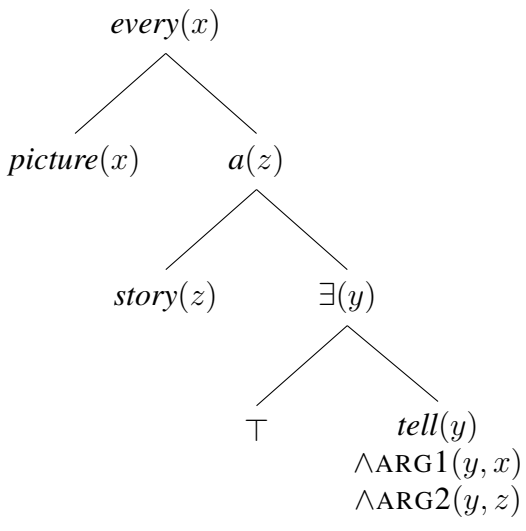

Figure 7: A scope tree representing the most likely reading of the sentence *every picture tells a story*. Each non-terminal node is a generalised quantifier, with its bound variable given in brackets. Its left child is its restriction, and its right child is its body. Going up through the tree, one variable is quantified over at a time, until the root of the tree has no free variables.

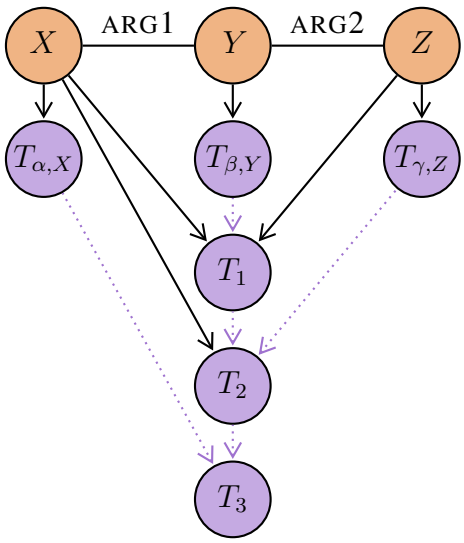

Figure 8: A probabilistic scope tree. T_1, T_2, T_3 correspond to non-terminal nodes in Fig. 7, going up through the tree. Solid lines indicate conditional dependence (on the *value* of the parent node), and dotted lines indicate the scope true (using the *distribution* of the parent node). One random variable is marginalised out at a time, until T_3 is no longer dependent on any variables.

Chapter 8: Conclusion

In the final chapter, I reflect on the achievements of the thesis, and give an outlook on future work. The main contribution of this thesis is a framework for distributional semantics which is compatible with model theory. Developing this framework has required extending model theory, and in particular, taking *generalisation* seriously. Furthermore, I have shown that a probabilistic model is more than just a computational convenience, but in fact allows a natural account of context dependence, which I have successfully applied to real-world datasets. I have also shown how the framework has a clear logical interpretation, and have presented a probabilistic generalisation of first-order logic with generalised quantifiers, which furthermore gives a natural account of vague quantifiers.

However, this thesis also represents the basis of a larger research project. Because the framework developed in this thesis is interpretable in linguistic terms and in logical terms, there is a clear and plausible path from this work to more general models of language. I outline four extensions, aiming towards the goals discussed in Chapter 2.

First, I discuss joint learning from multiple data sources, including both grounded data (such as images) and ontologies (such as WordNet). Second, I discuss semi-compositional expressions – for example, a *magic carpet* is both magic and a carpet, but is also able to fly. Semi-compositional constructions are poorly understood, both theoretically and computationally (for example: Sag et al., 2002; Reddy et al., 2011; Vincze, 2012). I sketch an account, building on the approach of Chapter 7. Third, I discuss pragmatics, which encompasses context dependence in a larger sense. I suggest that combining my framework with RSA (discussed in Chapter 7) would both extend my framework to include pragmatic reasoning, and allow us to scale up RSA to a large domain. Fourth, I discuss amortised variational inference and graph convolutions, two machine learning techniques which might significantly improve learning efficiency.

Reaching all of the goals given in Chapter 2 would be a breakthrough in computational linguistics. The framework I have developed in this thesis provides a basis from which we might hope to reach them.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of the 2009 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), Long Papers*, pages 19–27, 2009.

Keith Allan. *Natural language semantics*. Blackwell, 2001.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 39–48, 2016a.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Learning to compose neural networks for question answering. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 1545–1554, 2016b.

Antoine Arnauld and Pierre Nicole. *La logique ou l'art de penser*. Published by Jean Guignart, Charles Savreux, and Jean de Lavnay, 1662. Widely known as the *Port-Royal Logic*.

Marco Baroni, Raffaela Bernardi, and Roberto Zamparelli. Frege in space: A program of compositional distributional semantics. *Linguistic Issues in Language Technology (LiLT)*, 9. Center for the Study of Language and Information (CSLI) Publications, 2014.

Jon Barwise and Robin Cooper. Generalized quantifiers and natural language. *Linguistics and Philosophy*, 4(2):159–219. D. Reidel Publishing Company, 1981.

Jon Barwise and John Perry. *Situations and Attitudes*. Massachusetts Institute of Technology (MIT) Press, 1983.

Islam Beltagy, Stephen Roller, Pengxiang Cheng, Katrin Erk, and Raymond J. Mooney. Representing meaning with a combination of logical and distributional models. *Computational Linguistics*, 42(4):763–808. Massachusetts Institute of Technology (MIT) Press, 2016.

Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media Inc., 2009.

Georgeta Bordea, Els Lefever, and Paul Buitelaar. SemEval-2016 task 13: Taxonomy extraction evaluation (TExEval-2). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval)*, pages 1081–1091, 2016.

Arthur Bražinskas, Serhii Havrylov, and Ivan Titov. Embedding words as distributions with a Bayesian skip-gram model. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 1775–1789, 2018.

Elia Bruni, Giang Binh Tran, and Marco Baroni. Distributional semantics from text and images. In *Proceedings of GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 22–32, 2011.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. Multimodal distributional semantics. *Journal of Artificial Intelligence Research (JAIR)*, 49(2014): 1–47. 2014.

Jose Camacho-Collados, Claudio Delli Bovi, Luis Espinosa Anke, Sergio Oramas, Tommaso Pasini, Enrico Santus, Vered Shwartz, Roberto Navigli, and Horacio Saggion. SemEval-2018 task 9: Hypernym discovery. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval)*, pages 712–724, 2018.

Ronnie Cann. *Formal semantics: an introduction*. Cambridge University Press, 1993.

Gregory N. Carlson. *Reference to kinds in English*. PhD thesis, University of Massachusetts at Amherst, 1977.

Gregory N. Carlson and Francis Jeffry Pelletier, editors. *The generic book*. University of Chicago Press, 1995.

Rudolf Carnap. *Meaning and Necessity: A Study in Semantics and Modal Logic*. University of Chicago Press, 1947.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. Mathematical foundations for a compositional distributional model of meaning. *Linguistic Analysis*, 36, A Festschrift for Joachim Lambek:345–384. 2010.

Robin Cooper, Simon Dobnik, Staffan Larsson, and Shalom Lappin. Probabilistic type theory and natural language semantics. *Linguistic Issues in Language Technology (LiLT)*, 10. Center for the Study of Language and Information (CSLI) Publications, 2015.

Ann Copestake. Semantic composition with (Robust) Minimal Recursion Semantics. In *Proceedings of the ACL 2007 Workshop on Deep Linguistic Processing*, pages 73–80, 2007.

Ann Copestake. 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)*, pages 1–9, 2009.

Ann Copestake, Alex Lascarides, and Dan Flickinger. An algebra for semantic construction in constraint-based grammars. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 140–147, 2001.

Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A. Sag. Minimal Recursion Semantics: An introduction. *Research on Language and Computation*, 3(2-3):281–332. Springer, 2005.

Ann Copestake, Guy Emerson, Michael Wayne Goodman, Matic Horvat, Alexander Kuhnle, and Ewa Muszyńska. Resources for building applications with Dependency Minimal Recursion Semantics. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, pages 1240–1247, 2016.

Donald Davidson. The logical form of action sentences. In Nicholas Rescher, editor, *The Logic of Decision and Action*, pages 81–95. University of Pittsburgh Press, 1967. Reprinted in: Davidson (1980/2001), *Essays on Actions and Events*, Oxford University Press.

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

Katrin Erk. Supporting inferences in semantic space: representing words as regions. In *Proceedings of the 8th International Conference on Computational Semantics (IWCS)*, pages 104–115, 2009a.

Katrin Erk. Representing words as regions in vector space. In *Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL)*, pages 57–65, 2009b.

Katrin Erk. What is word meaning, really? (and how can distributional models help us describe it?). In *Proceedings of the 2010 Workshop on Geometrical Models of Natural Language Semantics*, pages 17–26, 2010.

Katrin Erk. What do you know about an alligator when you know the company it keeps? *Semantics and Pragmatics*, 9(17):1–63. 2016.

Manaal Faruqui, Yulia Tsvetkov, Dani Yogatama, Chris Dyer, and Noah Smith. Sparse overcomplete word vector representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL), Long Papers*, pages 1491–1500, 2015.

Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database* [Website]. Massachusetts Institute of Technology (MIT) Press, 1998.

Yansong Feng and Mirella Lapata. Visual information in semantic representation. In *Proceedings of the 2010 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 91–99, 2010.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited. In *Proceedings of the 10th International Conference on the World Wide Web*, pages 406–414, 2001.

John Rupert Firth. Modes of meaning. *Essays and Studies of the English Association*, 4:118–149. 1951. Reprinted in: Firth (1957), *Papers in Linguistics*, chapter 15, pages 190–215.

John Rupert Firth. A synopsis of linguistic theory 1930–1955. In John Rupert Firth, editor, *Studies in Linguistic Analysis*, Special volume of the Philological Society, pages 1–32. Blackwell, 1957.

Dan Flickinger, Stephan Oepen, and Gisle Ytrestøl. WikiWoods: Syntacto-semantic annotation for English Wikipedia. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, pages 1665–1671, 2010.

Michael C. Frank and Noah D. Goodman. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998. American Association for the Advancement of Science, 2012.

Peter Gärdenfors. *Conceptual spaces: The geometry of thought*. Massachusetts Institute of Technology (MIT) Press, 2000.

Peter Gärdenfors. *Geometry of meaning: Semantics based on conceptual spaces*. Massachusetts Institute of Technology (MIT) Press, 2014.

Dan Garrette, Katrin Erk, and Raymond Mooney. Integrating logical representations with probabilistic information using Markov logic. In *Proceedings of the 9th International Conference on Computational Semantics (IWCS)*, pages 105–114, 2011.

Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. SimVerb-3500: A large-scale evaluation set of verb similarity. In *Proceedings of the 21st Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2173–2182, 2016.

Noah D. Goodman and Michael C. Frank. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, 20(11):818–829. Elsevier, 2016.

Noah D. Goodman and Daniel Lassiter. Probabilistic semantics and pragmatics: Uncertainty in language and thought. In Shalom Lappin and Chris Fox, editors, *The Handbook of Contemporary Semantic Theory*, pages 655–686. Wiley, 2nd edition, 2015.

Edward Grefenstette and Mehrnoosh Sadrzadeh. Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of the 16th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1394–1404, 2011.

Edward Grefenstette, Georgiana Dinu, Yao-Zhong Zhang, Mehrnoosh Sadrzadeh, and Marco Baroni. Multi-step regression learning for compositional distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS)*, pages 131–142, 2013.

Patrick Hanks. Do word meanings exist? *Computers and the Humanities*, 34, Special Issue on SENSEVAL: Evaluating Word Sense Disambiguation Programs:205–215. Kluwer Academic Publishers, 2000.

Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42:335–346. Elsevier, 1990.

Zellig Sabbetai Harris. Distributional structure. *Word*, 10:146–162. Linguistic Circle of New York, 1954. Reprinted in: Harris (1970), *Papers in Structural and Transformational Linguistics*, chapter 36, pages 775–794; Harris (1981), *Papers on Syntax*, chapter 1, pages 3-22.

W. Keith Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109. Biometrika Trust, 1970.

Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING)*, pages 539–545, 1992.

Aurélie Herbelot. *Underspecified quantification*. PhD thesis, University of Cambridge, 2010.

Felix Hill, Roi Reichart, and Anna Korhonen. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695. Massachusetts Institute of Technology (MIT) Press, 2015.

Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800. Massachusetts Institute of Technology (MIT) Press, 2002.

Geoffrey E. Hinton. A practical guide to training Restricted Boltzmann Machines. Technical Report 2010-003, University of Toronto Machine Learning, 2010.

Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554. Massachusetts Institute of Technology (MIT) Press, 2006.

Manuela Hürlimann and Johan Bos. Combining lexical and spatial knowledge to predict spatial relations between objects in images. In *Proceedings of the 5th Workshop on Vision and Language*, pages 10–18, 2016.

Emily Elizabeth Constance Jones. *A New Law of Thought and its Logical Bearings*. Cambridge University Press, 1911.

Hans Kamp and Uwe Reyle. *From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory*, volume 42 of *Studies in Linguistics and Philosophy*. Springer, 2013.

David Kaplan. On the logic of demonstratives. *Journal of Philosophical Logic*, 8(1):81–98. Springer, 1979.

David Kaplan. Demonstratives: An essay on the semantics, logic, metaphysics, and epistemology of demonstratives and other indexicals. In Joseph Almog, John Perry, and Howard Wettstein, editors, *Themes from Kaplan*, pages 481–563. Oxford University Press, 1989.

Adam Kilgarriff. I don't believe in word senses. *Computers and the Humanities*, 31(2):91–113. Kluwer Academic Publishers, 1997.

Adam Kilgarriff. Word senses. In Eneko Agirre and Philip Edmonds, editors, *Word Sense Disambiguation: Algorithms and Applications*, pages 29–46. Springer, 2007.

Staffan Larsson. Formal semantics for perceptual classification. *Journal of Logic and Computation*, 25(2):335–369. Oxford University Press, 2013.

Daniel Lassiter. Vagueness as probabilistic linguistic knowledge. In Rick Nouwen, Robert van Rooij, Uli Sauerland, and Hans-Christian Schmitz, editors, *Vagueness in Communication: Revised Selected Papers from the 2009 International Workshop on Vagueness in Communication*, pages 127–150. Springer, 2011.

Daniel Lassiter and Noah D. Goodman. Adjectival vagueness in a Bayesian model of interpretation. *Synthese*, 194(10):3801–3836. Springer, 2015.

Angeliki Lazaridou, Elia Bruni, and Marco Baroni. Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world.

In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL), Long Papers*, pages 1403–1414, 2014.
Alessandro Lenci. Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics*, 20(1):1–31. 2008.
Alessandro Lenci. Distributional models of word meaning. *Annual Review of Linguistics*, 4:151–171. 2018.
Sarah-Jane Leslie. Generics: Cognition and acquisition. *Philosophical Review*, 117(1):1–47. Duke University Press, 2008.
Mike Lewis and Mark Steedman. Combined distributional and logical semantics. *Transactions of the Association for Computational Linguistics (TACL)*, 1:179–192. 2013.
Louise McNally. Kinds, descriptions of kinds, concepts, and distributions. In Kata Balogh and Wiebke Petersen, editors, *Bridging Formal and Conceptual Semantics: Selected Papers of BRIDGE-14*, pages 39–61. Düsseldorf University Press, 2017.
Louise McNally and Gemma Boleda. Conceptual versus referential affordance in concept composition. In *Compositionality and Concepts in Linguistics and Psychology*, number 3 in Language, Cognition, and Mind, pages 245–267. Springer, 2017.
Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092. American Institute of Physics, 1953.
Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of the 1st International Conference on Learning Representations (ICLR), Workshop Track*, 2013.
Ruth Garrett Millikan. Language conventions made simple. *The Journal of Philosophy*, 95(4):161–180. 1998. Reprinted in: Millikan (2005), *Language: A Biological Model*, chapter 1, pages 1–23, Oxford University Press.
Thomas P. Minka. Expectation propagation for approximate Bayesian inference. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*, pages 362–369, 2001.
Brian Murphy, Partha Pratim Talukdar, and Tom Mitchell. Learning effective and interpretable semantic models using non-negative sparse embedding. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, pages 1933–1950, 2012.
Charles K. Ogden and Ivor A. Richards. The meaning of meaning: A study of the influence of language upon thought and of the science of symbolism. Harcourt, Brace & World, Inc., 1923.
Terence Parsons. *Events in the Semantics of English: A Study in Subatomic Semantics*. Current Studies in Linguistics. Massachusetts Institute of Technology (MIT) Press, 1990.
Barbara H. Partee. Many quantifiers. In *Proceedings of the Eastern States Conference on Linguistics (ESCOL)*, pages 383–402, 1988. Reprinted in: Partee (2004), *Compositionality in Formal Semantics*, pages 241–258, Blackwell.
Behrang QasemiZadeh and Laura Kallmeyer. Random positive-only projections: PPMI-enabled incremental semantic space construction. In *Proceedings of the 5th Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 189–198, 2016.
Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. Massachusetts Institute of Technology (MIT) Press, 2006.
François Recanati. Compositionality, flexibility, and context-dependence. In Wolfram Hinzen, Edouard Machery, and Markus Werning, editors, *Oxford Handbook of Compositionality*, pages 175–191. Oxford University Press, 2012.
Siva Reddy, Diana McCarthy, and Suresh Manandhar. An empirical study on compositionality in compound nouns. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 210–218, 2011.
Laura Rimell, Jean Maillard, Tamara Polajnar, and Stephen Clark. RELPRON: A relative clause evaluation dataset for compositional distributional semantics. *Computational Linguistics*, 42(4):661–701. Massachusetts Institute of Technology (MIT) Press, 2016.
Eleanor Rosch. Cognitive representations of semantic categories. *Journal of experimental psychology: General*, 104(3):192. American Psychological Association, 1975.
Eleanor Rosch. Principles of categorization. In Eleanor Rosch and Barbara Bloom Lloyd, editors, *Cognition and categorization*, pages 27–48. Lawrence Erlbaum Associates, 1978. Reprinted in: Eric Margolis and Stephen Laurence, editors (1999), *Concepts: Core Readings*, chapter 8, pages 189–206.
Charles Ruhl. *On monosemy: A study in linguistic semantics*. State University of New York (SUNY) Press, 1989.
Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*, pages 1–15, 2002.
John R. Searle. The background of meaning. In John R. Searle, Ferenc Kiefer, and Manfred Bierwisch, editors, *Speech Act Theory and Pragmatics*, pages 221–232. D. Reidel Publishing Company, 1980.
Paul Smolensky. Information processing in dynamical systems: Foundations of harmony theory. In David E. Rumelhart and James L. McClelland, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume I: Foundations*, pages 194–281. Massachusetts Institute of Technology (MIT) Press, 1986.
Lars Jørgen Solberg. *A Corpus Builder for Wikipedia*. Master's thesis, University of Oslo, 2012.
Karen Spärck-Jones. *Synonymy and Semantic Classification*. PhD thesis, University of Cambridge, 1964. Reprinted in 1986 by Edinburgh University Press.
Peter R. Sutton. Towards a probabilistic semantics for vague adjectives. In Henk Zeevat and Hans-Christian Schmitz, editors, *Bayesian Natural Language Semantics and Pragmatics*, pages 221–246. Springer, 2015.
Peter R. Sutton. Probabilistic approaches to vagueness and semantic competency. *Erkenntnis*. Springer, 2017.
Kevin Swersky, Ilya Sutskever, Daniel Tarlow, Richard S. Zemel, Ruslan R. Salakhutdinov, and Ryan P. Adams. Cardinality Restricted Boltzmann Machines. In *Advances in Neural Information Processing Systems 25 (NIPS)*, pages 3293–3301, 2012.
Michael Henry Tessler and Noah D. Goodman. A pragmatic theory of generic language. Unpublished draft, 2016.
Peter D. Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37: 141–188. 2010.
Johan Van Benthem. Questions about quantifiers. *The Journal of Symbolic Logic*, 49(2):443–466. Association for Symbolic Logic, 1984.
Luke Vilnis and Andrew McCallum. Word representations via Gaussian embedding. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.
Veronika Vincze. *Semi-compositional noun + verb constructions: Theoretical questions and computational linguistic analyses*. PhD thesis, University of Szeged, 2012.
Ludwig Wittgenstein. *Philosophical Investigations*. Blackwell, 1953. Translated by Gertrude Elizabeth Margaret Anscombe. The original German text was published in 1958 under the title *Philosophische Untersuchungen*.
Kimberly Wong, Frempongma Wadee, Gali Ellenblum, and Michael McCloskey. The devil's in the g-tails: Deficient letter-shape knowledge and awareness despite massive visual experience. *Journal of Experimental Psychology: Human Perception and Performance*. American Psychological Association, 2018.
Jonathan S. Yedidia, William T. Freeman, and Yair Weiss. Understanding Belief Propagation and its generalizations. In Gerhard Lakemeyer and Bernhard Nebel, editors, *Exploring Artificial Intelligence in the New Millennium*, pages 239–269. Morgan Kaufmann Publishers, 2003.
Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics (TACL)*, 2:67–78. 2014.
Lotfi A. Zadeh. Fuzzy sets. *Information and Control*, 8(3):338–353. Academic Press, 1965.
Lotfi A. Zadeh. The concept of a linguistic variable and its application to approximate reasoning—I. *Information Sciences*, 8(3):199–249. Elsevier, 1975.