Probabilistic Semantics for Natural Language

Jan van Eijck and Shalom Lappin

CWI and ILLC Amsterdam, King's College London
jve@cwi.nl, shalom.lappin@kcl.ac.uk

Abstract

Probabilistic and stochastic methods have been fruitfully applied to a wide variety of problems in grammar induction, natural language processing, and cognitive modeling. In this paper we explore the possibility of developing a class of combinatorial semantic representations for natural languages that compute the semantic value of a (declarative) sentence as a probability value which expresses the likelihood of competent speakers of the language accepting the sentence as true in a given model, relative to a specification of the world. Such an approach to semantic representation treats the pervasive gradience of semantic properties as intrinsic to speakers' linguistic knowledge, rather than the result of the interference of performance factors in processing and interpretation. In order for this research program to succeed, it must solve three central problems. First, it needs to formulate a type system that computes the probability value of a sentence from the semantic values of its syntactic constituents. Second, it must incorporate a viable probabilistic logic into the representation of semantic knowledge in order to model meaning entailment. Finally, it must show how the specified class of semantic representations can be efficiently learned. We construct a probabilistic semantic fragment and consider how the approach that the fragment instantiates addresses each of these three issues.

1 Introduction

A formal semantic theory recursively defines the denotation of an expression in terms of the denotations of its syntactic constituents. It computes the semantic values of a sentence as a function of the values of its syntactic constituents. Within such a theory the meaning of an expression is identified with a function from indices (the expressions themselves, worlds, situations, times, etc.), to denotations in a model. The meaning of a sentence is a function from indices to truth-values.

Formal semantic theories model both lexical and phrasal meaning through categorical rules and algebraic systems that cannot accommodate gradience effects. This approach is common to theories which sustain compositionality and those with employ underspecified representations.[1] It effectively invokes the same strong version of the competence-performance distinction that categorical models of syntax assume. This view of linguistic knowledge has dominated linguistic theory for the past fifty years.

Gradient effects in representation are ubiquitous throughout linguistic and other cognitive domains. Appeal to performance factors to explain gradience has no explanatory content unless it is supported by a precise account of how the interaction of competence and performance generates these effects in each case. By contrast, gradience is intrinsic to the formal models that information theoretic methods use to represent events and processes.

Bach (1986) identifies two theses on the character of natural language.

(a) **Chomsky's thesis:** Natural languages can be described as formal systems.

(b) **Montague's thesis:** Natural languages can be described as *interpreted* formal systems.

Recent work in computational linguistics and cognitive modeling suggests a third proposal.

(c) **The Harris-Jelinek thesis:** Natural languages can be described as information theoretic systems, using stochastic models that express the distributional properties of their elements.

The Harris-Jelinek thesis implies the *The Language Model Hypothesis* (LMH) for syntax, which holds that grammatical knowledge is represented as a stochas-

[1] See, *inter alia*, Reyle (1993), Bos (1995), Blackburn and Bos (2005), Copestake et al. (2006), Koller et al. (2008), Fox and Lappin (2010) for discussions of underspecified semantics.

tic language model.[2] On the LMH, a speaker acquires a probability distribution $D : \Sigma^* \rightarrow [0, 1]$, over the strings $s \in \Sigma^*$, where Σ is a set of words (morphemes, etc.) of the language, and $\sum p_D(s) = 1$. This distribution is generated by a probabilistic automaton or a probabilistic grammar, which assigns a structure to a string with a probability that is the product of the rules applied in the derivation of that string. The probability of the string itself is the sum of the parses that the grammar generates for it. This probability represents the likelihood of a sentence's occurrence in a corpus.[3]

Representing linguistic knowledge stochastically does not eliminate the competence – performance distinction. It is still necessary to distinguish between a probabilistic grammar or automaton that generates a language model, and the parsing algorithm that implements it. However, a probabilistic characterization of linguistic knowledge does alter the nature of this distinction. The gradience of linguistic judgements and the defeasibility of grammatical constraints are now intrinsic to linguistic competence, rather than distorting factors contributed by performance mechanisms.

Lexically mediated relations like synonymy, antinomy, polysemy, and hyponymy are notoriously prone to clustering and overlap effects. They hold for pairs of expressions over a continuum of degrees [0,1], rather than Boolean values {1,0}. Moreover, the denotations of major semantic types, like the predicates corresponding to CNs, AdjPs, and VPs, can rarely, if ever, be identified as sets with determinate membership. The case for abandoning the categorical view of competence and adopting a probabilistic model is at least as strong in

[2]See Clark and Lappin (2011) for a discussions of the merits and problems of the LMH. An obvious difficulty with the LMH is that in the primary linguistic data for language acquisition short, ill formed sentences consisting of high frequency lexical items may have higher probability than longer, complex, well formed sentences containing low frequency words. A possible solution to this problem is to model grammatical acceptability in stochastic terms by imposing a lower bound on the probability of an acceptable string s that is dependent on properties of s, like its length, and features of the distribution for Σ^*. So, for example, a three word string like *You is here* is likely to have lower probability than the average probability of three word strings consisting of the word class sequence $\langle N, V, ADV \rangle$. By contrast, the string *Trading in complex instruments like mortgage backed derivatives and credit default swaps remains opaque and inexplicably under-regulated, which continues to be a major cause of instability in the financial markets* can be expected to have at least the average probability of strings of the same length and word class sequence. This approach to modeling acceptability uses the idea that one's expectation for the likelihood of occurrence of a string in a corpus depends, in part, on its properties and those of the distribution for its string set. It is derived from the stochastic model of indirect negative evidence that Clark and Lappin (2011) propose.

[3]See Manning and Schütze (1999), Collins (2003), Jurafsky and Martin (2009), Chelba (2010), Clark and Lappin (2010), Clark (2010) for discussions of statistical parsing and probabilistic grammars.

semantics as it is in syntax (as well as in other parts of the grammar)

A probabilistic semantics needs to express the probability of a different property than occurrence in a corpus. Knowing the meaning of a declarative sentence consists largely in being able to estimate the probability that competent speakers of the language would take it to be true across different states of the world (different worlds). This view is a probabilistic extension of a classical truth-conditional view of meaning. It can be extended to non-declarative sentences by formulating fulfillment conditions for them and identifying the meaning of a sentence with the function that determines the probability that speakers of the language construe it as fulfilled (a question answered, an imperative obeyed, a request satisfied, etc.) in any given state of affairs.[4]

As in the case of parsing, adopting a probabilistic view of semantic knowledge does not entail the eradication of the distinction between competence and performance. We still need to separate the semantic representation that generates a probability distribution for sentences in relation to states of affairs from the application of this representation in interpreting sentences. But like probabilistic grammars, these models incorporate gradience as an intrinsic feature of the objects that they characterize.

In this paper we argue that by replacing truth-conditions with probability conditions we can capture at least some of the pervasive gradience effects in semantic judgements. This allows us to reduce a number of important varieties of vagueness to the sort of uncertainty of belief (in this case, semantic belief) that probabilistic theories are designed to model. We are also able to account for several important kinds of semantic learning as a process of updating a learner's probability distributions over the worlds (which encode possible knowledge states) in which he/she evaluates the predicates whose meanings he/she is acquiring. This approach is consistent with the Harris-Jelinek thesis in that it represents semantic knowledge as a probability distribution over worlds that is generated by a probabilistic model for interpreting expressions of a language.

In Section 2 we present definitions of a model, a basic type theory, and a recursive definition of an interpretation function for a fragment of a formal representation language. In Section 3 we propose the outline of an account of semantic learning in which learners acquire the interpretation of new predicates, treated as probabilistic classifiers, in their language. We compare our approach to distributional treatments of meaning, particularly vector space models (VSMs), in Section 4. VSMs have emerged as highly efficient proce-

[4] Lappin (1982) offers an early proposal for characterizing truth conditions as an instance of fulfillment conditions.

dures for learning semantic relations among lexical items in corpora. Recent work has focussed on extending these methods to sentences. We discuss the complex connections among probability, gradience, and semantic vagueness in Section 5. Finally, in Section 6 we draw conclusions from our proposals and indicate directions for future work.

2 Probabilistic Models for a Semantic Fragment

Classical probabilistic logic Carnap (1950), Nilsson (1986), Fagin and Halpern (1991), Paris (2010) models uncertainty in our knowledge of the facts about the world. Probability distributions are specified over a set of possible states of the world (possible worlds), and the probabilities for the elements of this set sum to 1. A proposition ϕ is assigned truth-values across worlds, and ϕ's probability is computed as $\sum_{w \in W} p(w)$ for $\{w : \|\phi\|^w = t\}$.

In characterizing meaning probabilistically, we can talk of uncertainty about the truth-value of a sentence, given some probability distribution over possible states of affairs. The probability of a sentence expresses the likelihood that (semantically) competent speakers of the language assign to the truth of the sentence, given the state of their knowledge about the world. We can then represent the meaning of a sentence as a function that maps intensions to functions from knowledge states to probabilities (probability conditions). The semantic value of a sentence S is of type $I \to K \to [0,1]$, where I is the set of intensions, K is the set of knowledge representations, and $[0,1]$ is the set of reals p with $0 \leq p \leq 1$.

Let a propositional language over a set of basic predications be given, as follows.

$$\begin{array}{rcl} t & ::= & x \mid a_1 \mid a_2 \mid \cdots \mid a_m \\ Q & ::= & Q_1 \mid Q_2 \mid \cdots \mid Q_n \\ \phi & ::= & Qt \mid \neg\phi \mid \phi \wedge \phi \mid \phi \vee \phi. \end{array}$$

Here we assume a single variable x, a finite number of proper names $a_1, a_2, \ldots, a_m$ and a finite number of basic unary predicates $Q_1, Q_2, \ldots, Q_n$.

Any ϕ that contains occurrences of x is called a *predication*. Use $\phi(x)$ for predications, and $\phi(a/x)$ for the result of replacing x by a everywhere in a predication.

Call this language L_n^m. If we extend L_n^m with one name a_{m+1}, the new language is called L_n^{m+1}. If we extend L_n^m with one new predicate Q_{n+1}, the new language

is called L^m_{n+1}

For convenience, we identify names and objects, so we assume a domain $D_m = \{a_1, a_2, \ldots, a_m\}$. The type of a (restricted) world w is given by $w : \{Q_1, \ldots, Q_n\} \to \mathcal{P}(D_m)$. $w(Q_i)$ is the interpretation of Q_i in w.

A *probabilistic model* M is a tuple $\langle D, W, P\rangle$ with D a domain, W a set of worlds for that domain (predicate interpretations in that domain), and P a probability function over W, i.e., for all $w \in W$, $p(w) \in [0,1]$, and $\sum_{w\in W} p(w) = 1$.

An interpretation of L^m_n in an L^m_n-model $M = \langle D, W, P\rangle$ is given in terms of the standard notion $w \models \phi$, as follows:

$$[\![\phi]\!]^M \quad := \quad \sum\{P(w) \mid w \in W, w \models \phi\}$$

It is straightforward to verify that this yields $[\![\neg\phi]\!]^M = 1 - [\![\phi]\!]^M$. Also, if $\phi \models \neg\psi$, i.e., if $W_\phi \cap W_\psi = \emptyset$, then $[\![\phi \vee \psi]\!]^M = \sum_{w\in W_{\phi\vee\psi}} P(w) = \sum_{w\in W_\phi} P(w) + \sum_{w\in W_\psi} P(w) = [\![\phi]\!]^M + [\![\psi]\!]^M$, as required by the axioms of Kolmogorov (1950)'s probability calculus.

2.1 A Toy Fragment

Basic types are e (entities), s (worlds), t (truth values), d (domains) and $[0,1]$ (the space of probabilities). Abbreviate $d \to s \to t$ as i (intensions). Types for S, N, VP, NP, DET are lifted to the level of intensions, by substituting i for t in all types. This gives, e.g., $DET = (e \to i) \to (e \to i) \to i$.

The lifting rules for the interpretation functions are completely straightforward.

$I(Some) = \lambda p \lambda q \lambda dom \lambda w.\mathbf{some}(\lambda x.p\ x\ dom\ w)(\lambda y.q\ y\ dom\ w)$.

Here **some** is the familiar constant function for existential quantification, of type $(e \to t) \to (e \to t) \to t$.

This type system gives sentences an interpretation of type i, i.e., $d \to s \to t$. Such intensions can be mapped to probabilities by means of a function *prob* of type $i \to m \to [0,1]$, where m is the type of models with their domains, i.e., objects of the shape $\langle D, W, P\rangle$.

The function *prob* on sentences f and models $M = \langle D, W, P\rangle$ is given by:

$prob\ f\ \langle D, W, P\rangle = \sum\{P(w) \mid w \in W, f\ D\ w\}$

This function assigns to every sentence of the fragment a probability, on the basis of the prior probabilities encoded by $\langle D, W, P \rangle$.

2.2 Semantic Priors

The probabilities in a model M are the prior of a target semantic representation. We can take this prior to encode the knowledge representation that competent speakers converge upon as they acquire the meanings of the predicates of their language. Learners start out with different priors (probability distributions over models) than mature speakers, and update them through semantic learning. The prior that a learner brings to the learning task constitutes his/her initial assumptions about the state of the world, and, in a sense, it is the basis for semantic learning

Kemp et al. (2007) propose a hierarchical Bayesian learning framework in which observational classifiers and the learning priors that express expectations concerning the distribution of observations categorized by these classifiers can be acquired simultaneously from the same data. The priors are themselves derived from more general higher-order priors.

3 Semantic Learning

Classical semantic theories characterize a class of representations for the set of meanings of expressions in natural language. However, it is unclear how these representations could be learned from the primary linguistic data (PLD) of language acquisition. The problem of developing a plausible account of efficient learnability of appropriate target representations is as important for semantics as it is for other types of linguistic knowledge. Most work in formal learning for natural languages has focussed on syntax (grammar induction), morphology, and phonology.

3.1 Simple Cases of Learning

3.1.1 Example 1

Assume there are just two predicates Q_1 and Q_2, and two objects a, b. Complete ignorance about how the predicates are applied is represented by a model with 16 worlds, because for each object x and each predicate Q there are two cases: Q

applies to x or not. If the likelihood of each of the cases is completely unknown, each of these worlds has probability $\frac{1}{16}$.

3.1.2 Example 2

Suppose again there are two objects a, b and two predicates Q_1, Q_2. Assume that it is known that a has Q_1, and the probability that b has Q_1 is taken to be $\frac{2}{3}$. Suppose it is known that no object has Q_2. Then $W = \{w_1, w_2\}$ with $w_1(Q_1) = \{a, b\}$, $w_2(Q_1) = \{a\}$, $w_1(Q_2) = \emptyset$, $w_2(Q_2) = \emptyset$. P is given by $P(w_1) = \frac{2}{3}$, $P(w_2) = \frac{1}{3}$. In this example $\neg Q_1(b)$ is true in w_2 and not in w_1. Therefore $[\![\neg Q_1(b)]\!] = \frac{1}{3}$.

3.1.3 Learning New Definable Predicates

Learning a new semantic concept Q_{n+1} is learning how (or to what extent) predicate Q_{n+1} applies to the objects one knows about. The simplest way to model such a learning event is as a pair $\langle Q_{n+1}, \phi(x)\rangle$ where $\phi(x)$ is an L_n^m predication. The effect of the learning event could then be modeled in a way that is very similar to the manner in which factual change is modeled in an epistemic update logic.

The result of updating a model $M = \langle D, W, P\rangle$ with concept learning event $\langle Q_{n+1}, \phi(x)\rangle$ is the model that is like M except for the fact that the interpretation in each world of Q_{n+1} is given by

$$w(Q_{n+1}) := \{a \mid a \in D_m, w \models \phi(a/x)\}$$

Note that the probability function P of the model does not change in this case.

Let's return to example 1. This is the model where there are two objects and two predicates, and nothing is known about the properties of the objects. Take the learning event $(Q_3, Q_1x \wedge \neg Q_2x)$. This defines Q_3 as the difference of Q_1 and Q_2. The resulting model will again have 16 worlds, and in each world $w_i, w_i(Q_3)$ is given by $w_i(Q_1) \cap (D - w_i(Q_2))$. Again, the probabilities of the worlds remain unchanged.

3.2 Adjusting the Meaning of a Predicate

To allow adjustment of the meaning of a classifier by means of a learning event, we can use probabilistic updating (following **?**). A classifier learning event now

is a tuple $\langle Q, \phi, \psi(x), q\rangle$ where ϕ is a sentence, $\psi(x)$ is a predication, and q is a probability. ϕ expresses the observational circumstances of the revision. q expresses the observational certainty of the new information.

The result of updating $M = \langle D, W, P\rangle$ with $\langle Q, \phi, \psi(x), q\rangle$ is a new model $M = \langle D, W', P'\rangle$. W' is given by changing the interpretation of Q in members w of W_ϕ to $\{a \mid w \models \psi(a/x)\}$, while leaving the interpretation of Q in members of $W_{\neg\phi}$ unchanged.

P' is given by $P'(w) = \frac{P(w)\times q}{X}$ for members of W_ϕ, and by $P'(w) = \frac{P(w)\times(1-q)}{X}$ for members of $W_{\neg\phi}$. $\frac{1}{X}$ (the normalization factor) is given by

$$X = \sum_{w\in W_\phi} P(w)\times q + \sum_{w\in W_{\neg\phi}} P(w)\times(1-q).$$

3.2.1 Learning Classifiers by Example

Consider again the example with the two objects and the two properties, where new information concerning the application of the predicates to objects in the domain is acquired. A learning event for this could be $\langle Q_2, \neg Q_1 b, Q_1 x \vee Q_2 x, \frac{2}{3}\rangle$. Then the resulting model has again 2 worlds, but now the probability of w_2 has gone up from $\frac{1}{3}$ to $\frac{\frac{2}{3}\times\frac{1}{3}}{\frac{1}{3}\times\frac{2}{3}+\frac{2}{3}\times\frac{1}{3}} = \frac{1}{2}$. The probability of w_1 has gone down from $\frac{2}{3}$ to $\frac{\frac{1}{3}\times\frac{2}{3}}{\frac{4}{9}} = \frac{1}{2}$.

You are given something of which you are told that it is called a "rose", and you observe that it is thorny, red and a flower. A learning example is an encounter with a new object a_{m+1}. Suppose you learn that predicate Q applies to a_{m+1}. The properties you observe of a_{m+1} are given by $\theta(a_{m+1})$, where $\theta(a_{m+1})$ is a conjunction of $\pm Q_i(a_{m+1})$ for all known predicates. The update event is $\langle a_{m+1}, Q, \theta(a_{m+1})\rangle$. You learn that a_{m+1} is called a Q, and you observe that a_{m+1} satisfies the properties $\theta(a_{m+1})$.

Updating a model $M = \langle D, W, P\rangle$ for L_n^m with this event creates a new model $M' = \langle D \cup \{a_{m+1}\}, W', P\rangle$ for L_n^{m+1}. The new model has domain $\{a_1, \ldots, a_{m+1}\}$. W' is given by assigning, in each w, to a_{m+1} the properties specified by $\theta(a_{m+1})$. The interpretation of Q is given by setting $w(Q) = \{a \mid w \models \theta(a/a_{m+1})\}$. This resets the interpretation Q on the basis of the new observation. The probability distribution remains unchanged.

We can refine this account of learning to accommodate cases where an observation is less precise. Let the learning event be

$$\langle a_{m+1}, Q, \{(\theta_1(a_{m+1}), q_1), \ldots, (\theta_k(a_{m+1}), q_k)\}\rangle$$

Here q_i gives the observational probability that the new object satisfies θ_i. The probabilities should satisfy $\sum_{i=1}^{k} q_i = 1$. The update can be defined so that the probability of the new predicate applying to the old objects will be recomputed.

3.3 Semantic Knowledge and Knowledge of the World

Our specification of the class of probabilistic models and our treatment of learning raise the question of how to distinguish between semantic knowledge and knowledge of the world. It might seem that the distinction disappears entirely in our framework, and we are simply modeling epistemic update. In fact this is not the case. In a probabilistic account of epistemic update one seeks to express the effect of new information about the actual world on a belief agent's probability distribution over possible worlds. In our system of semantic representation we specify the meaning of a sentence as the likelihood that competent speakers of the language will assess it as true, given the distribution over worlds that sustains the interpretation of the expressions of their language. We are, then, seeking to model the probability that speakers assign to sentences across possible states of affairs, where these probability conditions are derived from the prior that speakers specify for worlds as a condition for sharing the meanings of their predicates. Semantic learning is a process of converging on the target model that generates this distribution by forming hypotheses on the intensions of predicates (the classifiers that they encode) on the basis of the PLD.

The notion of a semantic prior in terms of which the probability value of a sentence is computed allows us to identify semantic knowledge as distinct from general epistemic commitment. It is, however, the case that the distinction between semantic and extra-linguistic knowledge is not absolute. In learning a predicate one is acquiring a classifier that sorts objects on the basis of their properties. One could not apply such a classifier without recognizing these properties and making predictions concerning the likelihood that unobserved objects with similar properties satisfy (fail to satisfy) the classifier. It seems reasonable to assume that learners starting out with a semantic prior that is radically divergent from the target representation in most respects may find it difficult or impossible to acquire this representation from the PLD. If this does, in fact, turn out to be the case, then we can conclude that semantic learning depends on a core of shared beliefs about the nature of the world.

	context 1	context 2	context 3	context 4
financial	0	6	4	8
market	1	0	15	9
share	5	0	0	4
economic	0	1	26	12
chip	7	8	0	0
distributed	11	15	0	0
sequential	10	31	0	1
algorithm	14	22	2	1

Figure 1: Word Type-Context Matrix

4 Distributional Treatments of Meaning

4.1 Lexical Vector Space Models

Vector Space Models (VSMs) Turney and Pantel (2010) offer a fine-grained distributional method for identifying a range of semantic relations among words and phrases. They are constructed from matrices in which words are listed vertically on the left, and the environments in which they appear are given horizontally along the top. These environments specify the dimensions of the model, corresponding to words, phrases, documents, units of discourse, or any other objects for tracking the occurrence of words. They can also include data structures encoding extra-linguistic elements, like visual scenes and events.

The integers in the cells of the matrix give the frequency of the word in an environment. A vector for a word is the row of values across the dimension columns of the matrix. Figure 1 gives a schematic example of such a word-context matrix, with made up vector values. In this matrix the vectors for *chip* and *algorithm* are [7 8 0 0] and [14 22 2 1], respectively.

A pair of vectors from a matrix can be projected as lines from a common point on a plane. The smaller the angle between the lines, the greater the similarity of the terms, as measured by their co-occurrence across the dimensions of the matrix. Computing the *cosine* of this angle is a convenient way of measuring the angles between vector pairs. If $\vec{x} = \langle x_1, x_2, ..., x_n \rangle$ and $\vec{y} = \langle y_1, y_2, ..., y_n \rangle$ are two vectors, then

$$cos(\vec{x}, \vec{y}) = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \cdot \sum_{i=1}^{n} y_i^2}}$$

The cosine of $\vec{x}$ and $\vec{y}$ is their internal product, formed by summing the products of the corresponding elements of the two vectors, and normalizing the result relative to the lengths of the vectors. In computing $cos(\vec{x}, \vec{y})$ it may be desirable to apply a smoothing function to the raw frequency counts in each vector to compensate for sparse data, or to filter out the effects of high frequency terms. A higher value for $cos(\vec{x}, \vec{y})$ correlates with greater semantic relatedness of the terms associated with the $\vec{x}$ and $\vec{y}$ vectors.

VSMs provide highly successful methods for identifying a variety of lexical semantic relations, including synonymy, antinomy, polysemy, and hypernym classes. They also perform very well in unsupervised sense disambiguation tasks. VSMs offer a distributional view of lexical semantic learning. On this approach speakers acquire lexical meaning by estimating the environments (linguistic and non-linguistic) in which the words of their language appear.

4.2 Compositional VSMs

Lexical VSMs measure semantic distances and relations among words independently of syntactic structure. They apply a "bag of words" approach to meaning. Recent work has sought both to integrate syntactic information into the dimensions of the vector matrices Padó and Lapata (2007), and to extend VSM semantic spaces to the compositional meanings of sentences. Mitchell and Lapata (2008) compare additive and multiplicative models for computing the vectors of complex syntactic constituents, and they demonstrate better results (as measured by human annotator judgements) with the latter for sentential semantic similarity tasks. These models use simple functions for combining constituent vectors, and they do not represent the dependence of composite vectors on syntactic structure.

Coecke et al. (2010), Grefenstette et al. (2011) propose a procedure for computing vector values for sentences that specifies a correspondence between the vectors and the syntactic structures of their constituents. This procedure relies upon a category theoretic representation of the types of a pregroup grammar (PGG) Lambek (2008a;b), which builds up complex syntactic categories through direction-marked function application in a manner similar to a basic categorial grammar. All sentences receive vectors in the same vector space, and so they can be compared for semantic similarity using measures like cosine.

A PGG compositional VSM (CVSM) determines the values of a complex syntactic structure through a function that computes the tensor product of the vectors of its constituents, while encoding the correspondence between their grammatical types and their semantic vectors. For two (finite) vector spaces A, B, their tensor product $A \otimes B$ is constructed from the Cartesian product of the vectors in A and B. For any two vectors $v \in A$, $w \in B$, $v \otimes w$ is the vector consisting of all possible products $v_{i \in v} \times w_{j \in w}$. Smolensky (1990) uses tensor products of vector spaces to construct representations of complex structures (strings and trees) from the distributed variables and values of the units in a connectionist network.

PGGs are modeled as *compact closed categories*. A sentence vector is computed by a linear map f on the tensor product for the vectors of its main constituents, where f stores the type categorial structure of the string determined by its PGG representation. The vector for a sentence headed by a transitive verb, for example, is computed according to the equation

$$\overrightarrow{subj\ V_{tr}\ obj} = f(\overrightarrow{subj} \otimes \overrightarrow{V_{tr}} \otimes \overrightarrow{obj}).$$

The vector of a transitive verb V_{tr} could be taken to be an element of the tensor product of the vector spaces for the two noun bases corresponding to its possible subject and object arguments $\overrightarrow{V_{tr}} \in N \otimes N$. Then the vector for a sentence headed by a transitive verb could be computed as the point-wise product of the verb's vector, and the tensor product of its subject and its object

$$\overrightarrow{subj\ V_{tr}\ obj} = \overrightarrow{V_{tr}} \odot (\overrightarrow{subj} \otimes \overrightarrow{obj}).$$

PGG CVSMs offer a formally grounded and computationally efficient method for obtaining vectors for complex expressions from their syntactic constituents. They permit the same kind of measurement for relations of semantic similarity among sentences that lexical VSMs give for word pairs. They can be trained on a (PGG parsed) corpus, and their performance evaluated against human annotators' semantic judgements for phrases and sentences. Grefenstette and Sadrzadeh (2011) report that their system outperforms Mitchell and Lapata (2008)'s multiplicative CVSM in a small scale corpus experiment on predicting semantic distance for pairs of simple transitive VP sentences.

The PGG CVSM raises at least two major difficulties First, while the vector of a complex expression is the value of a linear map on the vectors of its parts, it is not obvious what independent property this vector represents. Sentential

vectors do not correspond to the distributional properties of these sentences, as the data in the primary linguistic data (PLD) from which children learn their language is too sparse to estimate distributional vectors for all but a few sentences, across most dimensions.

Coecke et al. (2010) show that it is possible to encode a classical model theoretic semantics in their system by using vectors to express sets, relations, and truth-values. But this simply demonstrates the formal power of PGG CVSMs as semantic coding devices. CVSMs are empirically interesting to the extent that the sentential vectors that they assign are derived from lexical vectors that represent the actual distributional properties of these expressions.

In classical formal semantic theories the functions that drive semantic composition are supplied by the type theory, where the type of each expression specifies the formal character of its denotation in a model. The sequence of functions that determines the semantic value of a sentence exhibits at each point a value that directly corresponds to an independently motivated semantic property of the expression to which it is assigned. Types of denotation provide non-arbitrary formal relations between types of expressions and classes of entities specified relative to a model. The sentential vectors obtained from distributional vectors of lexical items lack this sort of independent status. In our fragment we have specified a conservative extension of a classical type system for computing probabilistic values for sentences and predicates. An important advantage of our approach is that we sustain the independently motivated denotations that a classical type system assigns to syntactically complex expressions within a probabilistic framework designed to capture the gradience and relative uncertainty of lexical semantic relations.

The second major problem is as follows. An important part of the interpretation of a sentence involves knowing its truth (more generally, its satisfaction or fulfillment) conditions. We have exchanged truth conditions for probability conditions formulated in terms of the likelihood of a sentence being accepted by competent speakers of the language as true, given certain states of affairs in the world. It is not obvious how we can extract either classical truth conditions, expressed in Boolean terms, or probability conditions, from sentential vector values, when these are computed from vectors expressing the distributional properties of their constituent lexical items. By contrast, our fragment offers a recursive specification of the meaning of a sentence which yields its probability conditions.

5 Probability, Gradience, and Vagueness

5.1 Two Views of Semantic Vagueness

The fact that sentences receive probability conditions that express the likelihood that competent speakers would accept them as true relative to states of affairs permits us to model the uncertainty that characterizes some of these speakers' judgements concerning the semantic relations and predications that hold for their language. This sort of uncertainty accounts for an important element of gradience in semantic knowledge. It captures the defeasibility of implications, and the graded nature of synonymy (co-intensionality) and meaning intersection. However, it remains unclear whether all species of semantic vagueness can be subsumed by the uncertainty that probabilistic judgements express. Consider, in particular, the case of degree adjectives and adverbs. If a door is slightly ajar, there is a sense in which it fully satisfies neither *open* nor *closed*.[5]

Two views (*inter alia*) have been proposed for determining the relation between probability and semantic vagueness. On one of these, vagueness can be characterized in terms of the truth of judgements that predicates apply to objects, modifiers to states or events, etc. The epistemicist account of vagueness Williamson (1994) provides a prominent instance of this approach. It takes vagueness to consist in the same sort of uncertainty that attaches to epistemic claims about the world. This view is attractive to the extent that it can be used to support the idea that one models the gradience of semantic properties as a probability distribution over the applicability of expressions of different functional types to their arguments. However, it has the unattractive consequence that it assumes the existence of sharp boundaries on the extensions of predicates, but takes these to be epistemically opaque (essentially unknowable) to speakers of the language. Applying a predicate to an entity is, in many cases then, analogous to making a bet on the existence of a state of affairs, where one cannot identify the situation that decides the outcome of the wager. There appears to be no independent motivation for such unknowable limits on the extensions of terms. Therefore, it looks like an ad hoc device which the theory requires in order to explain the fact that vagueness, unlike epistemic uncertainty, cannot be eliminated by additional information about either language or the world.

[5] We are grateful to Peter Sutton for helpful discussion of the issues that we deal with in this section.

? offers a refined alternative version of the view that vagueness is the expression of probability judgements. He avoids the epistemicist assumption of unknowable determinate predicate extensions, by replacing these with a set of possible languages all of whose expressions receive non-vague interpretations. Vagueness is the result of a probability distribution over these languages (their predicates) in different worlds. Speakers assign probabilities to language-world pairs, seeking to maximize the probability of pairs that converge on the observed linguistic and non-linguistic facts. This analysis characterizes a vague predicate as ambiguous among a large disjunction of semantically determinate variants over which probability is distributed. In order to express the gradient nature of vagueness it would seem to be necessary to proliferate a large (possibly unbounded) number of determinate readings for vague predicates to range over. This looks like an awkward result. Vagueness is naturally thought of an alternative to ambiguity rather than a consequence of it.

? proposes the second view. She uses a Bayesian probability logic to model semantic vagueness, but she argues that vagueness and epistemic uncertainty are distinct. The problem with this approach is that it leaves the formal isomorphism between the two phenomena unexplained. If they really are different in the way that she suggests, then why should a calculus for computing the probability of statements under uncertainty provide a more accurate system for representing the vagueness of predicates than fuzzy or supervaluational logics, as she shows to be the case? The success of probabilistic models in expressing vagueness suggests that there is, in fact, a non-accidental connection between reasoning under conditions of epistemic uncertainty and the vagueness of predication. However, it may not be as direct or straightforward as the epistemicists hold it to be.

5.2 Semantic Vagueness as an Effect of Learning

It might be possible to develop a third view by mediating the relation between probability and vagueness through learning. Speakers learn predicates by generalizing from paradigm instances where their applications to an object are valued as 0 or 1 in worlds of high probability. Extending the application of these predicates to new objects with different property sets will produce an update in the probability function of the model that estimates the likelihood of competent speakers assenting to the predications as intermediary or low. In the absence of additional disambiguating evidence, this probability distribution over worlds for a range of predicate applications will survive learning to be incorporated into the model of mature speakers. In this way uncertainty in learning becomes

vagueness in the intensions of predicates in the target representation.

This approach treats epistemic uncertainty as a central element of semantic learning. The concern to converge on the classifiers that competent speakers apply drives the learner to update his/her probability distributions for the application of predicates (and other terms) in light of new linguistic and extra-linguistic evidence. But once the target representation is (more or less) achieved, many terms of the language remain under determined for objects in their domain. Vagueness is, then, the residue of probabilistic learning. It cannot be resolved by additional facts, linguistic or extra-linguistic, as it has been incorporated into the adult language itself. Therefore, it has its origin in probabilistic judgements on the truth of predication during the learning process, but it becomes an independent feature of the semantics of the language.

We offer this suggestion as the sketch of an alternative account of vagueness that seeks to account for it in probabilistic terms, but does not reduce it to epistemic uncertainty in the competent speakers of the language. In order to be viable it is necessary to work out a detailed formal theory of semantic learning and the target language that it converges on. This is a research project that this paper is intended to introduce, rather than complete.

6 Conclusions and Future Work

Compositional VSMs can represent gradience in semantic relations among words, phrases, and sentences, and they offer a viable account of lexical semantic learning. However, the vectors that CVSMs assign to complex syntactic structures do not have clear interpretations, and they do not express sentential meaning as probability conditions.

We propose a fragment of a probabilistic semantic theory that uses a conservative extension of classical type theory to compute the probability value of a sentence on the basis of a model for the knowledge of a semantic learner. Our approach offers a framework for developing a probabilistic account of semantic learning that is consonant with current Bayesian approaches to classifier acquisition.

We suggest a view of vagueness that treats it as originating in the probabilistic judgements of semantic learning, but which develops into an independent non-epistemic variety of uncertainty in the mature target representation language.

Acknowledgements The initial research for this paper was done when the

second author visited the first at the CWI for a month in the summer of 2011. The second author expresses his gratitude to the CWI for its generous hospitality and for the stimulating working environment that it provided during this time. Earlier versions of this paper were presented to a joint ILLC colloquium of the Computational Linguistics group and the DIP at the University of Amsterdam in September 2011, the King's College London Philosophy Colloquium in December 2011, The Oxford Advanced Seminar on Informatic Structures in December 2011, and the Hebrew University Logic, Language, and Cognition Center Colloquium in January 2012. We thank the participants of these meetings for their useful feedback and helpful suggestions. The second author is also grateful to Peter Sutton for helpful comments on an earlier draft of this paper, and for illuminating discussions of the relation between probability and semantic vagueness. These discussions have caused him to develop and refine some of the ideas proposed here. Of course we bear sole responsibility for any errors in the paper.

References

E. Bach. Natural languate metaphysics. In R. Barcan-Marcus, G. Dorn, and P. Weingartner, editors, *Logic, Methodology, and Philosophy of Science*, volume VII, pages 573-595. North Holland, Amsterdam, 1986.

P. Blackburn and J. Bos. *Representation and Inference for Natural Language*. CSLI, Stanford, 2005.

J. Bos. Predicate logic unplugged. In *Proceedings of the Tenth Amsterdam Colloquium*. Amsterdam, Holland, 1995.

R. Carnap. *Logical Foundations of Probability*. Uniiversity of Chicago Press, Chicago, 1950.

C. Chelba. Statistical language modeling. In A. Clark, C. Fox, and S. Lappin, editors, *Handbook of Computational Linguistics and Natural Language Processing*, pages 74–104. Wiley-Blackwell, Chichester, West Sussex and Malden, MA, 2010.

A. Clark and S. Lappin. Unsupervised learning and grammar induction. In A. Clark, C. Fox, and S. Lappin, editors, *Handbook of Computational Linguistics and Natural Language Processing*. Wiley-Blackwell, Oxford, 2010.

A. Clark and S. Lappin. *Linguistic Nativism and the Poverty of the Stimulus*. Wiley-Blackwell, Chichester, West Sussex, and Malden, MA, 2011.

S. Clark. Statistical parsing. In A. Clark, C. Fox, and S. Lappin, editors, *Handbook of Computational Linguistics and Natural Language Processing*, pages 333–363. Wiley-Blackwell, Chichester, West Sussex and Malden, MA, 2010.

B. Coecke, M. Sadrzadeh, and S. Clark. Mathematical foundations for a compositional distributional model of meaning. *Linguistic Analysis, Festschrift for Joachim Lambek*, 36:345–384, 2010.

M. Collins. Head-driven statistical models for natural language parsing. *Computational Linguistics*, 29(4):589–637, 2003.

A. Copestake, D. Flickinger, C. Pollard, and I. Sag. Minimal recursion semantics. *Research on Language and Computation*, 3:281–332, 2006.

R. Fagin and J. Halpern. Uncertainty, belief, and probability. *Computational Intelligence*, 7:160–173, 1991.

C. Fox and S. Lappin. Expressiveness and complexity in underspecified semantics. *Linguistic Analysis, Festschrift for Joachim Lambek*, 36:385–417, 2010.

E. Grefenstette and M. Sadrzadeh. Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1394–1404, Edinburgh, Scotland, 2011.

E. Grefenstette, M. Sadrzadeh, S. Clark, B. Coecke, and S. Pulman. Concrete sentence spaces for compositional distributional models of meaning. In *Proceedings of the 9th International Conference on Computational Semantics (IWCS-11)*, pages 125–134, Oxford, UK, 2011.

D. Jurafsky and J. Martin. *Speech and Language Processing*. Second Edition, Prentice Hall, Upper Saddle River, NJ, 2009.

C. Kemp, A. Perfors, and J. Tenenbaum. Learning overhypotheses with hierarchical bayesian models. *Developmental Science*, 103:307–317, 2007.

A. Koller, M. Regneri, and S. Thater. Regular tree grammars as a formalism for scope underspecification. In *Proceedings the 46th Annual Meeting of the ACL*. Columbus, OH, 2008.

A. Kolmogorov. *Foundations of Probability*. Chelsea Publishing, New York, 1950.

J. Lambek. Pregroup grammars and chomsky's earliest examples. *Journal of Logic, Language and Information*, 17(2):141–160, 2008a.

J. Lambek. *From Word to Sentence*. Polimetrica, Milan, 2008b.

S. Lappin. On the pragmatics of mood. *Linguistics and Philosophy*, 4:559–578, 1982.

C. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 1999.

J. Mitchell and M. Lapata. Vector-based models of semantic composition. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics 2008*, pages 236–244, 2008.

N. Nilsson. Probabilisitic logic. *Artificial Intelligence*, 28:71–87, 1986.

S. Padó and M. Lapata. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):177–199, 2007.

J. Paris. Pure inductive logic. Winter School in Logic, Guangzhou, China, 2010.

U. Reyle. Dealing with ambiguities by underspecification: Construction, representation and deduction. *Journal of Semantics*, 10:123–179, 1993.

P. Smolensky. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46:159–216, 1990.

P. Turney and P. Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188, 2010.

T. Williamson. *Vagueness*. Routledge, London, 1994.