From compositional semantics to Bayesian pragmatics via logical inference

Julian Grove **Jean-Philippe Bernardy** **Stergios Chatzikyriakidis**
Centre for Linguistic Theory and Studies in Probability
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg
firstname.lastname@gu.se

Abstract

Formal semantics in the Montagovian tradition provides precise meaning characterisations, but usually without a formal theory of the pragmatics of contextual parameters and their sensitivity to background knowledge. Meanwhile, formal pragmatic theories make explicit predictions about meaning in context, but generally without a well-defined compositional semantics. We propose a combined framework for the semantic and pragmatic interpretation of sentences in the face of probabilistic knowledge. We do so by (1) extending a Montagovian interpretation scheme to generate a distribution over possible meanings, and (2) generating a posterior for this distribution using a variant of the Rational Speech Act (RSA) models, but generalised to arbitrary propositions. These aspects of our framework are tied together by evaluating entailment under probabilistic uncertainty.[1]

We apply our model to anaphora resolution and show that it provides expected biases under suitable assumptions about the distributions of lexical and world-knowledge. Further, we observe that the model's output is robust to variations in its parameters within reasonable ranges.

1 Introduction

A goal of much work in computational semantics is to determine how responsibility should be apportioned between discrete, logical techniques and stochastic, probabilistic ones in explanations of inference. A current tradition that has roots in symbolic AI leverages the power of theorem provers to model inference in corpora, oftentimes grappling with both deductive and abductive modes of reasoning (Blackburn and Bos, 2005; Bos and Markert, 2005; Raina et al., 2005; van Eijck and Unger, 2010; Abzianidze, 2015; Emerson and Copestake, 2017a,b; Abzianidze, 2020, i.a.). Such approaches, while explicitly compositional, often attempt to combine both semantic and pragmatic meaning into a single inferential module, with the goal of capturing naturally occurring patterns.

Simultaneously (in the last decade), Rational Speech Act (RSA) models have provided a promising avenue for integrating logical and probabilistic approaches to meaning by modelling utterance interpretation as a process of updating probability distributions over logically characterised meanings (Goodman and Stuhlmüller, 2013; Lassiter and Goodman, 2013; Goodman and Frank, 2016; Lassiter and Goodman, 2017, i.a.). According to the RSA perspective, interpreting an utterance involves reasoning pragmatically about a speaker's intended message according to Bayesian principles of belief update. The reasoning of rational conversation participants, moreover, reflects principles of cooperative communication according to which speakers make true and informative utterances. Thus such models aim to capture a central feature of rational discourse known since the work of Grice (1975): that it is constrained by principles of appropriate social behaviour, which, through the reasoning of interlocutors, serve to enrich the very meanings which are communicated.

The goal of the current work is to integrate these two approaches to meaning and inference by using, on the one hand, a theorem prover to reason about compositionally derived semantic meanings and, on the other hand, Bayesian inference, as applied within the RSA framework, to give a computational account of pragmatic reasoning in discourse. Our contribution is thus to tie work in the logical tradition into a successful probabilistic framework for pragmatic reasoning. While logical entailment is at the core of evaluating truth values, we use probabilistic reasoning to deal with epistemic un-

[1] The code for this paper is available on GitHub at: https://github.com/juliangrove/grove-bernardy-chatzikyriakidis-naloma2021

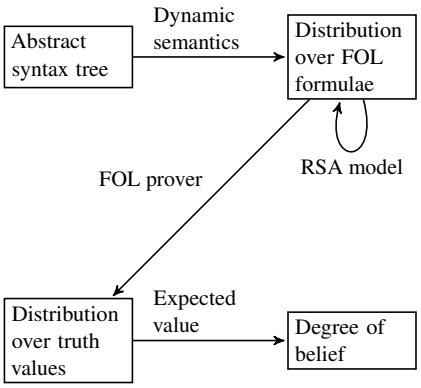

Figure 1: Phases of our system. Syntax is first interpreted into a distribution over FOL formulae. The truth values of such formulae can then be extracted using a theorem prover, allowing an expected value for the resulting distribution to be computed. The RSA model acts on the distribution over FOL formulae. This refinement step may itself invoke the theorem prover and distributions over truth values (we omit these dependencies to avoid clutter).

certainty. As such, this paper contributes a hybrid logical/probabilistic semantics consisting of both a standard Montagovian compositional scheme and an RSA model.

We pay special attention to the resolution of linguistic ambiguity in discourse—in particular, anaphora—as a test-case for our approach. By considering anaphora resolution as a Bayesian inference problem, we show how both prior world and lexical knowledge may influence the choice of antecedent for a given pronoun. Moreover, because our computational implementation combines a probabilistic approach to *inference* with a compositional, logical approach to *meaning* in the tradition of Montague (1973), i.e., by integrating numeric computation and theorem proving, we are able to explicitly and robustly characterise the contribution of conventional meaning to the task of pragmatic inference, as well as how the latter serves to modulate uncertainty about the former. We illustrate our approach on two test cases which differ in the priors they involve, thus demonstrating the importance of background knowledge to the behaviour of our model.

2 The framework

2.1 Compositional semantics under ambiguity

The goal of Montagovian compositional semantics is to map syntactic representations of an utterance u onto a meaning in some predefined domain. Typically, such meanings are propositions (for some logical system, like first order logic or type theory). We can write '$\phi = [\![u]\!]$' to represent such a mapping. However, there are, in general, several ways to map utterances to propositions, due to semantic ambiguity; thus our compositional semantics should instead produce a *distribution* of propositions.

The structure of our framework is illustrated in Figure 1. The first step in mapping an utterance to a pragmatically enriched meaning involves taking that utterance onto a probability distribution over expressions in some metalanguage: those which represent the dynamic semantic meaning of the utterance. As will become clear, any metalanguage for which one may define some computable notion of entailment suffices. For our implementation, we choose standard first order logic, so that utterances are mapped to distributions over FOL formulae.

Generalising the work of Lassiter and Goodman (2013), we may formalise this distribution in terms of the equation $\phi = [\![u]\!]^\theta$, where θ is a set of random variables, each having some *a priori* initial distribution. One can understand the above equation as invoking an interpretation function, $[\![\cdot]\!]$, defined inductively on expressions, and depending on the set θ of parameters whose role is to select among possible interpretations, as illustrated by the following scheme for Functional Application.

$$\frac{f = [\![np]\!]^\theta \quad x = [\![vp]\!]^\theta}{f(x) = [\![np\ vp]\!]^\theta}$$

In this way, a compositional semantics simultaneously produces distributions for ϕ and the parameters in θ.

2.2 Reasoning under probabilistic knowledge

One can evaluate the truth value of a proposition, given some background context, by evaluating its provability under a system of deduction (the system in question representing the reasoning capabilities of agents). We represent background knowledge as a distribution over sets of FOL formulae Γ, each of which may be regarded as representing a world-state; that is, a way things might be. We can then evaluate the truth value of ϕ, given some fixed world-state (i.e., set of hypotheses) Γ, as '$[\Gamma \vdash \phi]$', that is, 1 if $\Gamma \vdash \phi$ holds logically, and 0, otherwise. Even though entailment in many logical systems is undecidable, we may circumvent this issue, for

example, by limiting them to a certain depth of deduction, perhaps modelling finite reasoning capabilities. In our implementation, we use a regular FOL tableau prover limited to depth 10; this way, we can work with any set of propositions expressible in FOL, and in particular, those delivered by a Montagovian interpretation procedure. Calculating entailment constitutes the second step in our framework; it takes us from a probability distribution over formulae ϕ to a probability distribution over truth values reflecting whether ϕ holds at a world-state Γ, given an initial probability distribution over world-states.

Hereafter, we let Γ be a random variable ranging over sets of FOL formulae, whose distribution represents epistemic uncertainty of an agent about background knowledge, i.e., the actual world-state. Such a formulation of uncertainty is very flexible. For example, uncertainty about John's height can be represented as Γ = {John's height is H}, where H is a random variable with a normal distribution of mean 1.8 meters and standard deviation 0.05 meters. Discrete uncertainty may be represented using Bernoulli distributions. If b is a Boolean variable with a Bernoulli distribution, uncertainty about weather conditions can be represented as follows: Γ = if b then {it will rain tomorrow} else {it will not rain tomorrow}. Given a set of propositions $\Psi = \{\psi_1, \ldots, \psi_n\}$, each true or false according to one of a sequence of Bernoulli random variables $\gamma = b_1, \ldots, b_n$, we may take Γ to be equal to $\gamma\Psi$, i.e., the set containing either ψ_i or its negation $\neg\psi_i$, as according to whether b_i is True or False.

Given this setup, we can define a notion of 'expected truth value', which we encode as a real number between 0 and 1. We notate the truth value of ϕ, given some set of background hypotheses Γ, as '$[\Gamma \vdash \phi]$' and thus denote the expected truth value of ϕ, which takes into account the distribution associated with Γ, '$\mathbb{E}_\Gamma[\Gamma \vdash \phi]$'. As discussed in the next section, we will more often invoke the probability of non-entailment, given as $\mathbb{E}_\Gamma[\Gamma \nvdash \phi]$, and which is equal to $1 - \mathbb{E}_\Gamma[\Gamma \vdash \phi]$.

In general, we compute probability distributions over formulae, world-states, and truth values compositionally, in terms of probabilistic programs. A probabilistic program that returns a value of type α is a function of type $(\alpha \to \mathbb{R}) \to \mathbb{R}$; that is, one which consumes probability density functions (PDFs), i.e., from values of type α to real numbers, in order to derive a real number. For example, a probabilistic program that returns values of type α from some finite list l with a uniform distribution is the function $\lambda f.\texttt{sum}(\texttt{map}\, f\, l)/(\texttt{length}\, l)$. Given a PDF f, this program computes its sum across the members of l and divides the result by the length of l, thus returning the mean. If α is itself $\mathbb{R}$, the program may be used to compute an expected value by simply feeding it the identity function.

Crucially, probabilistic programs may be *composed*: given a probabilistic program m returning values of type $\alpha \to \beta$ and a probabilistic program n returning values of type α, a new program returning values of type β can be derived as $\lambda k.m(\lambda f.n(\lambda x.k(fx)))$. Such a composition scheme may appear familiar to many as applicative composition in the continuation monad. Indeed, probabilistic programs are composed by passing their input PDFs as continuations. More generally, complex probabilistic programs are easy to write and compose in monadic style, thus allowing us to keep our implementation pure and squarely within the simply typed λ-calculus. This approach, importantly, sets our framework apart from previous attempts at integrating natural language semantics with probabilistic computation, e.g., Goodman and Lassiter (2015).

2.3 RSA

We first review the general assumptions of the RSA model, and then we present the particular variant of RSA that we use in this paper. We discuss the differences between our presentation and that of the original RSA model of Lassiter and Goodman (2013) in §6.2.

RSA assumes two agents, a listener L and a speaker S. S utters a declarative sentence u heard by L, without transmission error. The point of RSA is to model how, assuming Gricean cooperativeness between S and L, L should disambiguate among possible interpretations of u.

Our model is defined by the following relations:

$$P_{L_1}(\phi \mid u) \propto P_{S_1}(u \mid \phi) \times P(\phi)$$
$$P_{S_1}(u \mid \phi) \propto (P_{L_0}(\phi \mid u)/C(u))^\alpha$$
$$P_{L_0}(\phi \mid u) = \mathbb{E}_{\theta,\Gamma}[\Gamma, \phi, [\![u]\!]^\theta \nvdash \bot]$$

In the above, relations for P_{L_1} and P_{L_0} represent listener models. Their primary function is to yield a distribution over interpretations of a given utterance u as propositions ϕ. $P_{L_1}(\phi|u)$ corresponds to a Bayesian update to the probability of the proposition ϕ, given an observation of the utterance u.

Following Bayes' theorem, this conditional probability is determined by multiplying a *likelihood*, $P_{S_1}(u \mid \phi)$, by a *prior*, $P(\phi)$.

The likelihood is the probability that the pragmatic listener S will utter u, given an intention to communicate the proposition ϕ. In other words, the output of the model P_{S_1} is an estimate of the probability of S uttering u if S *means* ϕ. This estimate is, in turn, obtained by considering utterances u in proportion to how likely they are to skew the (literal) listener towards interpreting u as ϕ, while taking into account an intrinsic utterance cost $C(u)$. Furthermore, an exponent α is applied to model the tendency of S to behave rationally, i.e., by choosing utterances in view of L_0's tendencies in conjunction with utterance cost.

The prior probability of the proposition ϕ is determined, in part, by the distribution over θ and, in part, by the distribution over prior knowledge Γ. Thus we have that $P(\phi) \propto \mathbb{E}_\Gamma[\Gamma, \phi \not\vdash \bot] * P(\theta)$, where $\phi = [\![u]\!]^\theta$ (for some θ), and $P(\phi) = 0$, otherwise. Priors are thus assessed using a *non-contradiction* model of interpretation: intuitively, Γ describes a world state—a way things could be—and ϕ is accepted if it is compatible with the world-state Γ.

The literal listener L_0 similarly uses a non-contradiction model of interpretation. It rejects interpretations ϕ incompatible with the utterance, in proportion to the *a priori* distribution of meanings for u, namely $[\![u]\!]^\theta$.

A final point deserving mention is that, in general, the priors over Γ and θ need not be the same in P_{L_1} and P_{L_0}. In P_{L_1}, they are L's actual priors, while in P_{L_0} they are those that L believes that S believes L has. In what follows, we consider only those priors which constitute common ground knowledge, i.e., in which case they are equal.

3 Anaphora resolution as a case study

To apply the above theory to anaphora resolution, we let θ be a set of parameters that determine the mapping of anaphoric expressions to antecedents. (In our experiments, we will consider only pronominal anaphora.)

For example, if $u =$ 'he runs', then (singleton) θ could be taken in the set $\Theta = \{John, Bill, Bob\}$, if those three antecedents are available in the discourse context. The logical representation of the utterance is then $[\![u]\!]^\theta = run(\theta)$. The factor $P(\theta)$ might be used to give lower probabilities to antecedents further back in the discourse; a value for this prior might be estimated from psycholinguistic experimentation. For the sake of simplicity, we let $P(\theta)$ be uniform across Θ in what follows.

Alternative utterances Within the RSA framework, P_{S_1} gives the distribution over alternative utterances considered by the speaker to express ϕ. The set underlying this distribution, moreover, must be supplied by the modeller *a priori*. We determine this set as follows. First, the utterance observed by L is itself in this set. Moreover, for any utterance u in the set, and for any anaphor x present in u, we include in the set the alternative utterance u' just like u, but in which the anaphor is substituted by a noun phrase denoting the antecedent actually meant by S (given S's intention to communicate ϕ). That is, u' is less ambiguous than u. For example, when evaluating $P(\text{'he runs'}|run(bill))$, S considers both 'he runs' and 'Bill runs'. This set is important because it is used to normalise S's distribution over utterances:

$$P_{S_1}(u|\phi) = \frac{(P_{L_0}(\phi|u)/C(u))^\alpha}{\sum_{u'}(P_{L_0}(\phi|u')/C(u'))^\alpha}$$

Background knowledge In our experiments, we let Γ be governed by a finite sequence of Boolean variables $b_1, \ldots, b_n$ drawn from Bernoulli distributions. Concretely, we work with a set of potential propositions $\{\psi_1, \ldots, \psi_n\}$ and write '$\ldots b_i \ldots \{\ldots \psi_i \ldots\}$' to denote the set containing ψ_i if b_i is True and $\neg\psi_i$ if b_i is False. The set of propositions in Γ and the parameters of their associated Bernoulli distributions vary from example to example.

The model then predicts a posterior distribution over mappings θ from anaphora to antecedents, along with a corresponding posterior distribution over meanings ϕ. As a result, one may also obtain a posterior distribution over Boolean variables b_i representing the common ground, now updated with ϕ.

4 Examples

We provide two examples to illustrate our model and, in particular, the effect of prior knowledge on its behaviour. Our first example is (1).

(1) Emacs is waiting for the command. It is prepared.

Here, the noun phrases *Emacs* and *the command* are in competition as potential antecedents for the

pronoun *it*.[2] Intuitively, the most likely antecedent for the pronoun is *Emacs*, which we take to be due (at least in part) to the fact that the verbs *waiting* and *prepared* lexically entail that their subjects are animate.[3] Thus a rational listener who infers that the antecedent for *it* in (1) is *Emacs* is doing so (at least in part) on the basis of the following reasoning: because animacy is entailed of the pronoun in virtue of its role as subject of the verb *prepared*, it is more likely, all else being equal, to co-refer with *Emacs*, which is also entailed to be animate (in virtue of being the subject of the verb *waiting*), than *the command*, which is subject to no such entailment. Such an inference is thus obtained on the basis of abductive reasoning about the source of the animacy of the pronoun.

The availability of this reasoning in (1) contrasts with its relative unavailibility in the second example in (2).

(2) Ashley is waiting for Amy. She sees her.

In contrast to *Emacs* and *the command*, proper names referring to humans, like *Ashley* and *Amy*, are very likely to denote animate individuals. As such, their prior probability of being animate will be higher than that of the noun phrases in (1), and the animacy entailment contributed by the verb *waiting* will therefore provide less of a basis for using animacy as a cue to distinguish potential antecedents for the pronouns. With the impact of animacy attenuated in (2), the candidate antecedents for the subject pronoun should be in closer competition, and anaphora resolution should be less certain. Intuitively, this seems to be the case: it appears more difficult to determine the referent of the subject pronoun in (2) than in (1) (though experimental investigation would be required to confirm this intuition).

We can model the difference between these examples by assuming different priors for the animacy of the referents of noun phrases like *Emacs* and *the command*, on the one hand, and *Ashley* and *Amy*, on the other. In our model, we encode such priors by associating probabilities with sentences translated into first order formulae; each such formula ψ is then associated with an independent Bernoulli random variable b in the definition of a probabilistic program that returns a world-state consisting of a set of hypotheses encoded as logical formulae. That is, such a set contains ψ if b is True, and it contains $\neg\psi$ if b is False.

$animate(emacs)$	0.2
$animate(the\_command)$	0.2
$animate(ashley)$	0.9
$animate(amy)$	0.9

As the table shows, we model world knowledge as dictating that the referents of *Emacs* and *the command* are only 20% likely to be animate, while individuals such as Ashley and Amy are 90% likely to be animate. Though these priors are somewhat arbitrary, they are meant to reflect qualitative differences in the knowledge we have about noun phrases referring to humans and those referring to other objects.

In addition to the priors listed above, we include priors for the truth of the following formulae, which, in each case, we take to be 0.05.

$$\exists x : wait\_for(emacs, x)$$
$$\exists x : wait\_for(the\_command, x)$$
$$\exists x : wait\_for(ashley, x)$$
$$\exists x : wait\_for(amy, x)$$
$$prepared(emacs)$$
$$prepared(the\_command)$$
$$\exists x : see(ashley, x)$$
$$\exists x : see(amy, x)$$

Finally, we encode the lexical entailments of the verbs *waiting*, *prepared*, and *sees* in terms of the following formulae:

$$\forall x : (\exists y : wait\_for(x, y)) \rightarrow animate(x)$$
$$\forall x : prepared(x) \rightarrow animate(x)$$
$$\forall x : (\exists y : see(x, y)) \rightarrow animate(x)$$

In our model, these formulae act as filters of background knowledge: any world-state that contradicts them is given probability 0, and the probability distribution over world-states is re-normalised. As a result, the Bernoulli random variables associated with individual hypotheses in the final model of background knowledge will not be entirely independent.

5 Results and analysis

To illustrate the model's performance, we give results for the examples discussed in the previous section, fixing values for parameters in the speaker model; in particular, the exponent α, as well as the log-cost associated with an utterance that uses either a pronoun or a full noun phrase to refer to a

[2] This example comes from Lappin and Leass (1994), who resolve anaphora on the basis of a number syntactic and semantic heuristics, with no specific pragmatic analysis.

[3] These inferences may, in fact, be presuppositions, a point we gloss over here.

given antecedent. Table 1 provides the model's calculations of the pragmatic listener's bias to choose *Emacs* (as opposed to *the command*) as the antecedent of the subject pronoun of (1), across two values of α and two sets of values for log-cost. Log-costs for pronouns (PN) and full noun phrases (NP) are summed, for any given utterance, to provide its total log-cost. For example, if the log-cost of a pronoun is 1, and that of a full noun phrase is 2 (as in the models reported in rows 2 and 4), then an utterance with one pronoun and one noun phrase will have a total log-cost of 3, and the probability P_{L_0} is scaled by a factor of $e^{-3\alpha}$ in the calculation of P_{S_1}.

α	PN	NP	*Emacs* bias
0.5	0	0	87.9%
0.5	1	2	86.9%
4.0	0	0	99.9%
4.0	1	2	98.6%

Table 1: Example (1)

The results of Table 1 highlight three notable features of our model. First, anaphora resolution displays the expected bias, based on the prior world and lexical knowledge governing inference. In particular, lexical knowledge associated with the verbs *waiting* and *prepared* determines that their subjects be animate; thus the pragmatic listener performs a kind of abductive inference, based on these entailments: a pronoun which is entailed to be animate displays a high probability of seeking animacy in its antecedent. Comparison with the results for (2) (which we discuss next) illustrates the importance of the low animacy priors (0.2) for the antecedents in achieving pragmatic reasoning of this kind.

Second, even though high values of α increase the bias in favour of *Emacs* (as expected), the model is not very sensitive to its precise choice. As α approaches 0, the speaker model approaches a uniform distribution over utterances, but even as low a value as 0.5 yields sensible results.

Third, incorporating a measure of cost into the reasoning of the pragmatic speaker has a dampening effect on the model's bias, as can be seen by comparing rows 1 and 2, as well as rows 3 and 4. This effect consists in about 1% of difference, and it is due to the fact that making reference to cost has the pragmatic listener reason about a "lazier" pragmatic speaker; such a speaker, who finds full noun phrases costlier to utter than pronouns, will more likely choose a pronoun to *minimise their effort*, rather than as a result of their reasoning about a literal listener who will choose the expected antecedent for the pronoun.

Table 2 provides the model's calculations of the pragmatic listener's bias to choose *Ashley* (as opposed to *Amy*) as the antecedent for both the subject pronoun *she* and the object pronoun *her* in (2). We show results for the same values of α and log-cost.

α	PN	NP	*Ashley* bias	
			for *she*	for *her*
0.5	0	0	53.0%	50%
0.5	1	2	52.9%	50%
4.0	0	0	60.7%	50%
4.0	1	2	54.2%	50%

Table 2: Example (2)

We note, first, that the same general patterns across values of α and log-cost obtain for this example as for the previous one: higher values of α exaggerate the pragmatic listener's bias, while increasing noun phrase cost relative to pronoun cost dampens it.

Second, comparing the results of this model with those for (1) demonstrates clearly the effect of prior knowledge on the model's behaviour. Because the antecedents have high animacy priors (0.9), the animacy entailment of the verb *waiting* provides less of a basis for distinguishing them; as a result, they are in closer competition as antecedents for the subject pronoun, which is entailed to be animate, and bias toward the subject antecedent is greatly reduced (though still present).

Last, we note that the object pronoun is exactly split in its probability of taking *Ashley* versus *Amy* as its antecedent in (2). Because there is no animacy entailment from the verb for the object pronoun, the pragmatic listener has no basis for distinguishing the antecedents, e.g., through abductive inference. This result supports our explanation for the biases displayed in the other cases.

6 Related work

The work presented in this paper is related to a number of attempts in both the formal and computational semantics communities to bridge logical and probabilistic approaches to natural language semantics. These approaches, in addition to their formal differences, can be categorised into those which have been computationally implemented and

those which have not. In the first category, one finds approaches such as Beltagy et al. (2013); Goodman and Stuhlmüller (2013); Goodman and Frank (2016); Lassiter and Goodman (2013, 2017); Bernardy et al. (2018); Emerson and Copestake (2017b) , while in the latter category, those such as van Eijck and Lappin (2012); Cooper et al. (2015); Sutton (2018).

A common theme among probabilistic approaches to interpretation is that they describe a set of possible world-states as a distribution. Predicates are then evaluated at each world-state, and probabilistic truth is the expected value over all possible world-states. In implemented accounts, one often uses Monte Carlo sampling methods to estimate truth values. We refrain from a further comparison with approaches lacking a computational implementation: even though they contain fruitful ideas, it is unclear how they should be realised computationally.

Another way to classify approaches is by the representation of world-states that they employ. Goodman and Stuhlmüller (2013); Goodman and Frank (2016); Lassiter and Goodman (2013) use an *ad hoc* set of variables, chosen according to the problem at hand. Bernardy et al. (2018) use vector representations inspired by machine-learning approaches. Bernardy et al. (2019b) present a system that tries to minimise (and in cases, eliminate) the need for sampling by modelling predicates as (the unions of) boxes and individuals as points.

A unique characteristic of the present account is our use of a small number of Bernoulli random variables to represent world-states, where each variable captures the applicability of a proposition. This choice is afforded by the use of logical entailment as the basis of evaluating truth values. Together, this means that we can provide exact calculations for truth values, i.e., by taking the average over finite probability distributions. An additional benefit of using the knowledge-as-propositions approach is that we have all the expressivity of the underlying logic at our disposal. Hence, we have no difficulty dealing with predicates with multiple arguments, contrary to Bernardy et al. (2019b, 2018). Even though weighted formulae can be interpreted as possible world-states via a Markov Logic Networks (Domingos and Lowd, 2009), as Beltagy et al. (2013) showed for natural language semantics, our simpler approach is sufficient for our purposes.

6.1 Logical approaches to semantic inference

Our framework aspires to connect two traditions in the study and computational implementation of semantics: logical, compositional semantics on the one hand, and Bayesian pragmatics, on the other. This connection is achieved by reasoning about propositional entailment via theorem proving, while modelling pragmatic inference as Bayesian reasoning, using a variant of RSA. Thus there are important connections to other approaches to semantics and natural language inference that rely on a compositional semantics to translate abstract syntax trees into logical formulae and then evaluate inference patterns via theorem proving (Bos and Markert, 2005; Mineshima et al., 2015; Abzianidze, 2015; Bernardy and Chatzikyriakidis, 2017, 2019, 2021). These accounts vary in their details; for example, in the type of parser used: Bos and Markert (2005); Mineshima et al. (2015); Abzianidze (2015) use variants of CCG parsers, while Bernardy and Chatzikyriakidis (2017, 2019) use the GF parser (Ranta, 2011). They also vary in the types of meaning representations they employ, as well as in the underlying logical systems they use (e.g., first order versus higher order). Finally, they differ in their choice of theorem provers, and whether they are automated or interactive. But the connections between such approaches and ours are clear: all employ a compositional semantics to generate logical formulae, which are further reasoned about with theorem provers. A crucial difference between our approach and the aforementioned ones, however, is that ours supports a designated pragmatic module that accomplishes pragmatic inference with Bayesian reasoning. Thus our framework may be seen as involving a pragmatic enrichment of a logical component, afforded by Bayesian reasoning in the guise of RSA. Finally, despite the fact that our account follows previous work in the RSA tradition (Goodman and Stuhlmüller, 2013; Goodman and Frank, 2016; Lassiter and Goodman, 2013, 2017), it employs a couple of different assumptions than usual—a point to which we now turn.

6.2 The relation between our model and standard RSA

Lassiter and Goodman (2013) give an RSA model governed by the following relations (modulo re-

naming of some parameters):

$$P_{L_1}(\Gamma, \theta \mid u) \propto P_{S_1}(u \mid \Gamma, \theta) \times P_{L_1}(\Gamma)$$
$$P_{S_1}(u \mid \Gamma, \theta) \propto (P_{L_0}(\Gamma \mid u, \theta)/C(u))^{\alpha}$$
$$P_{L_0}(\Gamma \mid u, \theta) = P_{L_0}(\Gamma \mid [\![u]\!]^{\theta})$$

This model differs from ours in two notable ways. First, the model of Lassiter and Goodman directly marginalises the distribution of world-states (Γ in the above formalisation), while we only consider the possible meanings of an utterance ($[\![u]\!]^{\theta}$). In other words, we regard pragmatic inference as a problem of inferring utterance meanings, rather than one of directly updating the common ground.

This choice has practical consequences from a modelling perspective. When applying the framework of Lassiter and Goodman, one needs to choose prior distributions carefully, in order to cover all possible aspects of a given world-state which may be relevant to the truth value of any of the possible meanings of u; i.e., those, which, in our example of anaphora resolution, we obtained as mappings from utterances to propositions that varied along the set of parameters θ.

Second, we have allowed the pragmatic speaker model P_{S_1} to marginalise over θ. In contrast, the model of Lassiter and Goodman uses a value of θ which is fixed throughout the model; i.e., it is passed up from the literal listener to the pragmatic listener. Since our distribution over θ depends on the utterance whose interpretation it parameterises, we allow our pragmatic speaker to re-sample θ in its model of the literal listener.

Finally, in comparison to previous RSA work which attempts to combine a natural language semantics with probabilistic reasoning (see Goodman and Lassiter, 2015), the approach we advocate is, we believe, conservative, flexible, and modular:

- It allows for the usual approach to compositional semantics, i.e., in a pure logical language.
- Any such logic can be chosen, so long as it is equipped with a computable notion of entailment.
- Probabilistic computation is added in terms of continuation passing, i.e., as a monadic side effect.
- Even such a side effect does not extend the basic semantics of the metalanguage, which is just the simply typed λ-calculus. We therefore end up with a compositional mathematical theory of the phenomena under investigation.

This situation contrasts, for example, with the implementation of Goodman and Lassiter (2015), using Church. While Goodman and Lassiter are innovative in their integration of probabilistic computation into a functional language, they extend the simply typed λ-calculus with a probabilistic semantics, which, as far as we can tell, is not entirely compositional and thus difficult to reason about.

7 Future directions

The model presented in this paper relies on techniques that are widely used in computational semantics; by combining them in a novel way, we believe that our approach has important potential to generate applications in semantic analysis, inference, and Bayesian cognitive modelling. One immediate avenue for extending our model concerns its applicability to a range of semantic problems that could benefit from a system that leverages both logical semantics and Bayesian reasoning. An obvious candidate is predication vagueness, a classic problem for logical semantics and the target of discussion of a number of Bayesian approaches to semantics (Sutton and Filip, 2016; Lassiter and Goodman, 2017; Bernardy et al., 2019a; Emerson, 2020). Thus extending the coverage of the present model and checking the predictions it makes with respect to these phenomena is one of our goals.

We are also interested in designing a general natural language inference system based on the approach proposed in this paper; such a system could then be evaluated against various test suites. To start, one can check whether the proposed system accounts for pragmatic aspects of the FraCaS test suite (Cooper et al., 1996), the RTE test suite (Dagan et al., 2006), or the small probabilistic test suite of Bernardy et al. (2019a).

As a realistic anaphora resolution algorithm, the model presented here falls short in some respects. First, we take no account of the well-studied grammatical restrictions on the relation between pronominal anaphora and their antecedents (Reinhart, 1976; Chomsky, 1981). Second, our model currently shows no sensitivity to the discourse factors which are well known to affect the acceptability of anaphora in various contexts. And third, it incorporates no sensitivity to psycholinguistic constraints on anaphora, which, like discourse factors,

affect acceptability. There are different ways one might make the model sensitive to such constraints, which may be decided on a case-by-case basis. In principle, any non-pragmatic factor may be accounted for by imposing the right prior on θ. However, other solutions suggest themselves. Grammatical constraints on the anaphora-antecedent relation, for example, might be implemented in an improved compositional semantics which makes antecedents available for certain anaphora depending on their relative syntactic positions. Sensitivity to discourse factors might be incorporated into our model as declarative knowledge that contributes to the prior (i.e., on a par with world and lexical knowledge). And, psycholinguistic (and, perhaps, discourse) constraints might, for example, be incorporated into our pragmatic speaker model as a more realistic measure of cost (see, e.g., Orita et al., 2015), or our literal listener model, by sampling antecedents for anaphora according to their retrieval costs. In principle, antecedent retrieval cost could be incorporated into the distribution over antecedents accessed by the pragmatic listener, as well, perhaps depending on whether our model is viewed as giving a computational-level versus algorithmic-level characterisation of anaphora resolution (Marr, 1982).

The ultimate success of our approach relies on obtaining an accurate account of prior knowledge. Prior world knowledge can be obtained through experiment, following approaches to RSA that have assessed prior beliefs using surveys (Xiang et al., 2021a,b). Given that our model characterises prior knowledge declaratively, we can use similar methods.

Finally, although we have paid specific attention to anaphora resolution, our model makes way for a general approach to semantic ambiguity resolution. We might, for example, extend our model to other anaphora-like phenomena, e.g., ellipsis, as well as the resolution of structural and quantifier-scope ambiguities. The success of such extensions depends on generating an appropriate set of alternatives, given an utterance (and vice versa). For ellipsis, semantic alternatives can be generated by a free parameter, as above; in the case of, e.g., quantifier scope, one might incorporate a parser to provide alternative semantic representations for a given utterance. In all cases, one requires an appropriate set of alternative utterances from a proposition in the speaker model.

Acknowledgements

The research reported in this paper was supported by grant 2014-39 from the Swedish Research Council, which funds the Centre for Linguistic Theory and Studies in Probability (CLASP) in the Department of Philosophy, Linguistics, and Theory of Science at the University of Gothenburg. We thank the anonymous reviewers for their useful comments on an earlier draft of the paper.

References

Abzianidze, L. (2015). A Tableau Prover for Natural Logic and Language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2492–2502, Lisbon, Portugal. Association for Computational Linguistics.

Abzianidze, L. (2020). Learning as Abduction: Trainable Natural Logic Theorem Prover for Natural Language Inference. *arXiv:2010.15909 [cs]*. arXiv: 2010.15909.

Beltagy, I., Chau, C., Boleda, G., Garrette, D., Erk, K., and Mooney, R. (2013). Montague Meets Markov: Deep Semantics with Probabilistic Logical Form. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 11–21, Atlanta, Georgia, USA. Association for Computational Linguistics.

Bernardy, J.-P., Blanck, R., Chatzikyriakidis, S., and Lappin, S. (2018). A Compositional Bayesian Semantics for Natural Language. In *Proceedings of the First International Workshop on Language Cognition and Computational Models*, pages 1–10, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Bernardy, J.-P., Blanck, R., Chatzikyriakidis, S., Lappin, S., and Maskharashvili, A. (2019a). Bayesian Inference Semantics: A Modelling System and A Test Suite. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 263–272, Minneapolis, Minnesota. Association for Computational Linguistics.

Bernardy, J.-P., Blanck, R., Chatzikyriakidis, S., Lappin, S., and Maskharashvili, A. (2019b). Predicates as Boxes in Bayesian Semantics for Natural Language. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 333–337, Turku, Finland. Linköping University Electronic Press.

Bernardy, J.-P. and Chatzikyriakidis, S. (2017). A Type-Theoretical system for the FraCaS test suite: Grammatical Framework meets Coq. In *IWCS 2017 - 12th International Conference on Computational Semantics - Long papers*.

Bernardy, J.-P. and Chatzikyriakidis, S. (2019). A Wide-Coverage Symbolic Natural Language Inference System. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 298–303, Turku, Finland. Linköping University Electronic Press.

Bernardy, J.-P. and Chatzikyriakidis, S. (2021). Applied temporal analysis: A complete run of the fracas test suite. In *IWCS 2021 - 14th International Conference on Computational Semantics - Long papers*.

Blackburn, P. and Bos, J. (2005). *Representation and Inference for Natural Language: A First Course in Computational Semantics*. Studies in Computational Linguistics. University of Chicago Press, Chicago.

Bos, J. and Markert, K. (2005). Recognising Textual Entailment with Logical Inference. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 628–635, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Chomsky, N. (1981). *Lectures on Government and Binding: The Pisa Lectures*. Number 9 in Studies in Generative Grammar. Foris Publications, Dordrecht.

Cooper, R., Crouch, D., Eijck, J. V., Fox, C., Genabith, J. V., Jaspars, J., Kamp, H., Milward, D., Pinkal, M., Poesio, M., Pulman, S., Briscoe, T., Maier, H., and Konrad, K. (1996). Using the Framework. Technical Report LRE 62-051 D-16, The FraCaS Consortium.

Cooper, R., Dobnik, S., Lappin, S., and Larsson, S. (2015). Probabilistic Type Theory and Natural Language Semantics. In *Linguistic Issues in Language Technology, Volume 10, 2015*. CSLI Publications.

Dagan, I., Glickman, O., and Magnini, B. (2006). The PASCAL Recognising Textual Entailment Challenge. In Quiñonero-Candela, J., Dagan, I., Magnini, B., and d'Alché Buc, F., editors, *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, Lecture Notes in Computer Science, pages 177–190, Berlin, Heidelberg. Springer.

Domingos, P. and Lowd, D. (2009). Markov Logic: An Interface Layer for Artificial Intelligence. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 3(1):1–155. Publisher: Morgan & Claypool Publishers.

Emerson, G. (2020). Linguists who use probabilistic models love them: Quantification in functional distributional semantics. In *Proceedings of the Probability and Meaning Conference (PaM 2020)*, pages 41–52, Gothenburg. Association for Computational Linguistics.

Emerson, G. and Copestake, A. (2017a). Semantic Composition via Probabilistic Model Theory. In *IWCS 2017 - 12th International Conference on Computational Semantics - Long papers*.

Emerson, G. and Copestake, A. (2017b). Variational Inference for Logical Inference. *arXiv:1709.00224 [cs]*. arXiv: 1709.00224.

Goodman, N. D. and Frank, M. C. (2016). Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences*, 20(11):818–829.

Goodman, N. D. and Lassiter, D. (2015). Probabilistic Semantics and Pragmatics Uncertainty in Language and Thought. In Lappin, S. and Fox, C., editors, *The Handbook of Contemporary Semantic Theory*, pages 655–686. John Wiley & Sons, Ltd. Section: 21 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118882139.ch21

Goodman, N. D. and Stuhlmüller, A. (2013). Knowledge and Implicature: Modeling Language Understanding as Social Cognition. *Topics in Cognitive Science*, 5(1):173–184.

Grice, H. P. (1975). Logic and Conversation. In Cole, P. and Morgan, J. L., editors, *Syntax and Semantics*, volume 3, Speech Acts, pages 41–58. Academic Press, New York.

Lappin, S. and Leass, H. J. (1994). An Algorithm for Pronominal Anaphora Resolution. *Computational Linguistics*, 20(4):535–561.

Lassiter, D. and Goodman, N. D. (2013). Context, scale structure, and statistics in the interpretation of positive-form adjectives. *Semantics and Linguistic Theory*, 23(0):587–610. Number: 0.

Lassiter, D. and Goodman, N. D. (2017). Adjectival vagueness in a Bayesian model of interpretation. *Synthese*, 194(10):3801–3836.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. MIT Press, Cambridge.

Mineshima, K., Martínez-Gómez, P., Miyao, Y., and Bekki, D. (2015). Higher-order logical inference with compositional semantics. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2061, Lisbon, Portugal. Association for Computational Linguistics.

Montague, R. (1973). The Proper Treatment of Quantification in Ordinary English. In Hintikka, K. J. J., Moravcsik, J. M. E., and Suppes, P., editors, *Approaches to Natural Language: Proceedings of the 1970 Stanford Workshop on Grammar and Semantics*, Synthese Library, pages 221–242. Springer Netherlands, Dordrecht.

Orita, N., Vornov, E., Feldman, N., and Daumé III, H. (2015). Why discourse affects speakers' choice of referring expressions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1639–1649, Beijing, China. Association for Computational Linguistics.

Raina, R., Ng, A. Y., and Manning, C. D. (2005). Robust textual inference via learning and abductive reasoning. In *Proceedings of the 20th national conference on Artificial intelligence - Volume 3*, AAAI'05, pages 1099–1105, Pittsburgh, Pennsylvania. AAAI Press.

Ranta, A. (2011). *Grammatical Framework: Programming with Multilingual Grammars*. CSLI Publications, Stanford.

Reinhart, T. M. (1976). *The syntactic domain of anaphora*. Thesis, Massachusetts Institute of Technology. Accepted: 2005-08-02T20:38:55Z.

Sutton, P. R. (2018). Probabilistic Approaches to Vagueness and Semantic Competency. *Erkenntnis*, 83(4):711–740.

Sutton, P. R. and Filip, H. (2016). Vagueness, Overlap, and Countability. *Proceedings of Sinn und Bedeutung*, 20:730–747.

van Eijck, J. and Lappin, S. (2012). Probabilistic semantics for natural language. In Christoff, Z., Galeazzi, P., Gierasimczuk, N., Marcoci, A., and Smets, S., editors, *Logic and interactive rationality (LIRA)*, volume 2, pages 17–35. Citeseer.

van Eijck, J. and Unger, C. (2010). *Computational Semantics with Functional Programming*. Cambridge University Press, Cambridge.

Xiang, M., Dai, Z., and Wang, S. (2021a). When Parsing and interpretation misalign: a case of wh-scope ambiguity resolution in Mandarin. Under review.

Xiang, M., Kennedy, C., Weijie, X., and Leffel, T. (2021b). Pragmatic Reasoning and Semantic Convention: A Case Study on Gradable Adjectives. Under review.