Pragmatic Reasoning through Semantic Inference

Leon Bergen, Roger Levy, and Noah D. Goodman

November 26, 2014

Abstract

A number of recent proposals have used techniques from game theory and Bayesian cognitive science to formalize Gricean pragmatic reasoning (Frank & Goodman, 2012; Franke, 2009; Goodman & Stuhlmüller, 2013; Jäger, 2012). We discuss two phenomena which pose a challenge to these accounts of pragmatics: M-implicatures (Horn, 1984) and embedded implicatures which violate Hurford's constraint (Chierchia, Fox & Spector, 2012; Hurford, 1974). Previous models cannot derive these implicatures, because of basic limitations in their architecture. In order to explain these phenomena, we propose a realignment of the division between semantic content and pragmatic content. Under this proposal, the semantic content of an utterance is not fixed independent of pragmatic inference; rather, pragmatic inference partially determines an utterance's semantic content. We show how semantic inference can be realized as an extension to the Rational Speech Acts framework (Goodman & Stuhlmüller, 2013). The addition of *lexical uncertainty* derives both M-implicatures and the relevant embedded implicatures, and preserves the derivations of more standard implicatures.

Keywords: Pragmatics, Game theory, Hurford's constraint, Embedded implicatures, Division of pragmatic labor, Bayesian modeling

1 Introduction

Theories of natural language semantics aim to provide a simple account of how people interpret expressions in their language. Attempts to provide such an account face a basic challenge: the interpretation of expressions frequently varies with linguistic and social context. An obvious response to such contextual variation is to posit that natural language expressions are highly polysemous. A naive implementation of this idea will have at least two deficiencies: the theory will need to be extremely complex to accommodate all of the possible meanings of each expression; and it will miss the systematic relationship between an expression's context and its interpretation.

Gricean theories of pragmatics provide an elegant solution to these problems. They posit that the interpretation of an expression is not necessarily identical to its semantic content. Rather, this semantic content plays a specific role in the derivation of the expression's interpretation. In typical circumstances, speakers and listeners regard each other as rational agents who share the goal of communicating information to each other. A speaker chooses an utterance by reasoning about the beliefs that a listener would form if they interpreted utterances according to their semantic content; the speaker will be more likely to choose an utterance that is effective at communicating their intended meaning. The listener, in turn, interprets an utterance by reasoning about which intended meanings would have made the speaker most likely to choose this utterance. Gricean pragmatic accounts thus factor the interpretation of an expression into two parts: its semantic content, which determines its literal meaning, and cooperative social reasoning, which builds on this literal interpretation to determine the expression's inferred meaning. By factoring out the role of semantic content in this manner, Gricean pragmatic accounts reduce the explanatory burden of semantic theories. Many facts about an expression's interpretation will be determined by the communicative setting in which the expression is used, and not simply the expression's semantic content.

Despite the promise and apparently broad empirical coverage of these theories, attempts at formalizing them (e.g., Gazdar, 1979) have historically met with less success than formalization in other linguistic domains such as phonology, syntax, or semantics. Nevertheless, there is strong reason to believe that formal accounts of Gricean pragmatic reasoning have substantial potential scientific value. First, all Gricean theories assume that multiple factors—most famously Grice's quality, quantity, relevance, and manner—jointly guide the flexible relationship between literal semantic content and understood meaning, and in all Gricean theories these factors can potentially come into conflict (e.g., the opposition between Horn's (1984) Q and R principles). Our success at cooperative communication implies that a calculus of how different factors' influence is resolved in each communicative act is broadly shared within every speech community, yet extant theories generally leave this calculus unspecified and are thus unsatisfactory in predicting preferred utterance interpretation when multiple factors come into conflict. Mathematical formalization can provide such a calculus. Second, in the decades since Grice's original work there has been a persistent drive toward conceptual unification of Grice's original maxims into a smaller set of principles (e.g., Horn, 1984; Levinson, 2000; Sperber & Wilson, 1986). Mathematical formalization can help rigorously evaluate which such efforts are sound, and may reveal new possibilities for unification. Third, the appropriate mathematical formalization may bring pragmatics into much closer contact with empirical data, by making clear (often quantitative) and falsifiable predictions regarding communicative behavior in specific situations that may be brought under experimental control. This kind of payoff from formalization has been seen in recent years in related fields including psycholinguistics (Lewis & Vasishth, 2005; Smith & Levy, 2013) and cognitive science (Tenenbaum, Kemp, Griffiths & Goodman, 2011). Fourth, the development of pragmatic theory necessarily has a tight relationship with that of semantic theory. A precise, formalized pragmatic theory may contribute to advances in semantic theory by revealing the nature of the literal meanings that are exposed to Gricean inference and minimizing the possibility that promissory appeals to pragmatics may leave key issues in semantics unresolved.

The last several years have, in fact, seen a number of recent accounts which formalize Gricean pragmatic reasoning using game theory or related utility-based decision-theoretic frameworks that are beginning to realize this potential (Degen, Franke & Jäger, 2013; Frank & Goodman, 2012; Franke, 2009; Franke & Jäger, 2013; Goodman & Stuhlmüller, 2013; Jäger, 2012; Parikh, 2000; Rothschild, 2013). These accounts find conceptual unification in grounding cooperative communicative behavior in simple principles of efficient information exchange by rational agents that can reason about each other. These accounts provide a precise specification of the reasoning that leads conversational partners to infer conversational implicatures either by using the notion of a game-theoretic equilibrium to define conditions that the agents' reasoning must meet or by providing a computational or procedural description of the reasoning itself. They characteristically provide formal proposals of the division between semantic content and pragmatic inference in which the semantic content of each linguistic expression is determined outside of the model, by a separate semantic theory. This semantic content serves as input to the pragmatics model, which in turn, specifies how agents use this semantic content, in addition to facts about their conversational setting, in order to infer enriched pragmatic interpretations of the expressions. Finally, by bringing in linking assumptions regarding the relationship between probabilistic beliefs and action from mathematical psychology, some of these models have been tested against empirical data far more rigorously than has been seen in previous work (Degen et al., 2013; Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013).

This paper continues these efforts, using recursive probabilistic models to formalize Gricean explanations of a sequence of increasingly complex pragmatic phenomena. We will begin by providing an account, in line with previous game-theoretic models, of scalar implicatures and generalized class of these implicatures, which we refer to as *specificity implicatures*. We will also demonstrate how this *rational speech acts model* provides a solution to the symmetry problem for scalar implicatures.

We will next turn to M-implicatures, inferences that assign marked interpretations to complex expressions. We will show that the simple model of specificity implicatures does not derive M-implicatures, for reasons that are closely related to a well-known problem in game theory, the multiple equilibrium problem for signalling games. In order to derive even the simplest types of M-implicatures, we need to relax the traditional Gricean factorization of semantic content and pragmatic inference. In particular, the semantic content of expressions will not be determined in advance of pragmatic inference. Rather, the participants in a conversation will jointly infer this semantic content, as they are performing pragmatic reasoning.

Semantic inference plays an essential role in the derivation of M-implicatures. In order to represent the speaker and listener's inferences about the semantic content of their language's expressions, we will introduce *lexical uncertainty*, according to which the speaker and listener begin their pragmatic reasoning uncertain about how their language's lexicon maps expressions to literal meanings. An important part of pragmatic reasoning thus involves resolving this semantic uncertainty, in addition to determining the non-literal content of the expressions. By extending the rational speech acts model with lexical uncertainty, we will be able to derive simple M-implicatures, in which complex expressions are assigned low probability interpretations. We will be able to derive a larger class of M-implicatures, in which complex utterances are assigned more generally marked interpretations, by relaxing the assumption that the speaker is knowledgeable.

Finally, we will consider a difficult class of embedded implicatures, which have not yet been derived within game-theoretic models of pragmatics. These implicatures cannot be derived by the rational speech acts model or the simple extension of this model with lexical uncertainty. In order to derive these implicatures, our model will need to be sensitive to the compositional structure of the expressions that it is interpreting. We will extend the model so that it respects the compositional structure of expressions, and represents uncertainty about the semantic content of genuine elements of the lexicon — i.e., atomic expressions — rather than whole expressions. When the model is extended in this manner, it will derive the embedded implicatures in question.

Though the determination of semantic content cannot be separated from pragmatic reasoning under our proposal — indeed, semantic inference will drive the derivation of the more interesting implicatures that we will consider — we will not have to abandon all of the explanatory advantages that factored Gricean accounts provide. Under our proposal, the explanatory burden of semantic theories will still be limited: they will need to account for approximately the same interpretive phenomena as they do under more traditional Gricean theories. As we will describe in more detail below, this is because the semantic content provided by semantic theories will still only play a limited functional role within our models. Our models primarily depart from traditional Gricean theories in their account of what role this semantic content will play.

2 The baseline rational speech-act theory of pragmatics

We begin by introducing the baseline rational speech-act theory of pragmatics (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013), built on a number of simple foundational assumptions about speakers and listeners in cooperative communicative contexts. We assume first a notion of COMMON KNOWLEDGE (Clark, 1996; Lewis, 1969; Stalnaker, 1978)—information known by both speaker and listener, with this shared knowledge jointly known by both speaker and listener, knowledge of the knowledge of shared knowledge jointly known by both speaker and listener, and so on *ad infinitum* (or at least as many levels of recursion up as necessary in the recursive pragmatic inference). Communication involves the transmission of knowledge which is not common knowledge: we assume that the speaker, by virtue of some observation that she has made, is in a particular belief state regarding the likely state of some conversationally relevant aspect of the world (or, more tersely, regarding the world). In engaging in a cooperative communicative act, the speaker and listener have the joint goal of bringing the listener's belief state as close as possible to that of the speaker, by means of the speaker formulating and sending a not-too-costly signal to the listener, who interprets it. The lexicon and grammar of the speaker and listener's language serve as resources by which literal content can be formulated. As pragmatically sophisticated agents, the speaker and the listener recursively model each other's expected production decisions and inferences in comprehension.

More formally, let O be the set of possible speaker observations, $\mathcal{W}$ the set of possible worlds, and $\mathcal{U}$ the set of possible utterances. Observations $o \in O$ and worlds $w \in \mathcal{W}$ have joint prior distribution $P(o,w)$, shared by listener and speaker. The literal meaning of each utterance $u \in \mathcal{U}$ is defined by a lexicon $\mathcal{L}$, which is a mapping from each possible utterance-world pair to the truth value of the utterance in that world. That is,

$$\mathcal{L}(u,w) = \begin{cases} 0 & \text{if } w \notin [\![u]\!] \\ 1 & \text{if } w \in [\![u]\!] \end{cases} \tag{1}$$

where $[\![u]\!]$ is the intension of u.[1] The first and simplest component of the model is the LITERAL LISTENER, who interprets speaker utterance u by conditioning on it being true and computing via Bayesian inference a belief state about speaker observation state o and world w. This updated distribution L_0 on w is defined by:

$$L_0(o,w|u,\mathcal{L}) \propto \mathcal{L}(u,w)P(o,w). \tag{2}$$

Social reasoning enters the model through a pair of recursive formulas that describe how the speaker and listener reason about each other at increasing levels of sophistication. We begin with

[1] Note that this definition of the lexicon departs from standard usage, as it assigns meanings to whole utterances rather than atomic subexpressions. This is a provisional assumption which will be revised in Section 5.

the speaker, who plans a choice of utterance based on the EXPECTED UTILITY of each utterance, with utterances being high in utility insofar as they bring the listener's belief distribution about world and speaker observation close to that of the speaker, and low in utility insofar as they are costly to produce. Discrepancy in belief distribution is measured by the standard information theoretic quantity of KULLBACK-LEIBLER DIVERGENCE from the listener's posterior distribution $L(o,w|u)$ on observation and world given utterance, to the speaker's distribution given observation $P(o,w|o)$, which reduces to $P(w|o)$ and we denote more compactly as P_o. The Kullback-Leibler divergence from distribution Q to distribution P is in general defined as

$$D_{\mathrm{KL}}(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}. \tag{3}$$

Therefore the discrepancy in speaker and listener belief distributions is quantified as

$$D_{\mathrm{KL}}(P_o||L^u) = \sum_w P(w|o) \log \frac{P(w|o)}{L(o,w|u)}. \tag{4}$$

where L^u is shorthand for $L(\cdot|u)$. We define the expected utility of utterance u for a recursion-level n speaker who has observed o as

$$U_n(u|o) = -D_{\mathrm{KL}}(P_o||L_{n-1}) - c(u) \tag{5}$$

where $c(u)$ is the cost of utterance u. Intuitively, utterances are costly insofar as they are time-consuming or effortful to produce; in this paper, we remain largely agnostic about precisely what determines utterance cost, assuming only that utterance cost is strictly monotonic in utterance lengths (as measured in words).

In the first part of this paper, we assume that for each world $w \in \mathcal{W}$, there is a unique observation $o \in O$ consistent with this world. In this special case, it is common knowledge that the speaker knows the true world w with probability 1, so that $P(w|o)$ is 1 for that world and 0 for all other worlds. This entails that we can ignore the world variable w in the speaker and listener equations, and the discrepancy in speaker and listener beliefs reduces to the negative log-probability or SURPRISAL of the observation for the listener given the utterance:

$$D_{\mathrm{KL}}(P_o||L^u) = \log \frac{1}{L(o|u)} = -\log L(o|u). \tag{6}$$

Under these conditions, (expected) utterance utility can be written as simply

$$U_n(u|o) = \log L_{n-1}(o|u) - c(u) \tag{7}$$

The assumption of speaker knowledgeability is relaxed in Section 4.6.

We are now ready to state the speaker's formula. The speaker's conditional distribution over utterances given the world w under consideration as the listener's possible interpretation is defined as

$$S_n(u|o) \propto e^{\lambda U_n(u|o)}, \tag{8}$$

where $\lambda > 0$. This specification of the speaker formula uses the SOFTMAX FUNCTION or LUCE-CHOICE RULE (Sutton & Barto, 1998) to map from a set of utterance utilities to a probability distribution over utterance choice. The INVERSE-TEMPERATURE parameter λ governs the speaker's degree of "greedy rationality". When $\lambda = 1$, the probability that the speaker chooses utterance u is proportional to the exponentiated utility of u. As λ increases, the speaker's distribution over utterance choices becomes increasingly more strongly peaked toward utterances with high exponentiated utility. The Luce-choice rule is used extensively in psychology and cognitive science as a model of human decision-making, and in reinforcement learning in order design algorithms that balance maximizing behavior that is optimal in the short-run and exploratory behavior that is beneficial in the long-run (Sutton & Barto, 1998).

Finally, we turn to the listener's recursive formula for interpreting utterances by reasoning about likely speaker choices. The listener's higher-order interpretations are simply defined as

$$L_n(o,w|u) \propto P(o,w)S_n(u|o). \tag{9}$$

That is, the listener uses Bayes' rule to reconcile their prior expectations about world state to be described with their model of the speaker. Equations (2), (5), (8), and (9) constitute the heart of this basic model. Note the relationship between recursion levels of the speaker and listener in Equations (5): the first speaker S_1 reasons about the literal listener L_0, the first pragmatic listener L_1 reasons about S_1, the second speaker S_2 reasons about the first pragmatic listener L_1, and so forth. The model we present here generalizes the rational speech-act model presented in Goodman & Stuhlmüller (2013) by adding utterance costs and the possibility of recursion beyond S_1.

2.1 Auxiliary assumptions: alternative sets, but no lexical scales

As in much previous work in pragmatics (Gazdar, 1979; Grice, 1975; Horn, 1984; Levinson, 2000), our models of pragmatic reasoning will rely heavily the set of alternative utterances available to the speaker. That is, in deriving the implicatures for an utterance, our models will reason about why the speaker did not use the other utterances available to them. We will not be providing a general theory of the alternative utterances that are reasoned about during the course of pragmatic inference. Rather, as is done in most other work in pragmatics, we will posit the relevant set of utterances on a case-by-case basis. As is discussed below, however, there are certain cases for which our models require fewer restrictions on the set of alternatives than most other models. These examples will provide suggestive — though not decisive — evidence that no categorical restrictions need to be placed on the alternatives set within our models, i.e. that every grammatical sentence in a language can be considered as an alternative during pragmatic reasoning. The mechanisms by which this may be made possible are discussed below.

Our models' treatment of lexical scales will represent a larger departure from the norm. By a "scale," we are referring to a totally ordered set of lexical items which vary along a single dimension; a typical example is the set of lexical items <"some", "most", "all">, where each item (when used in a sentence) is logically stronger than all of the items that fall below it on the scale. Such scales play an important role in many theories of pragmatic reasoning, where they constrain the set of alternative utterances available to the speaker. In such theories, it is assumed that the set of alternative utterances can be totally ordered along a relevant dimension (e.g. along the dimension of informativeness for ordinary scalar implicatures), so that this set forms a scale. Our models will not use scales in order to derive pragmatic inferences. In certain cases, the set of alternatives used by the model will include multiple utterances which are logically equivalent to each other. In other cases, the set of alternatives will include utterances which are jointly logically inconsistent.

In general, the global constraints on the alternatives set which are described by scales will not be required by our models.

3 Specificity implicature in the baseline theory

To demonstrate the value of the baseline theory presented in Section 2, we show here how it accounts for a basic type of pragmatic inference: specificity implicatures, a generalization of scalar implicatures, in the case where it is common knowledge that the speaker knows the relevant world state. Specificity implicatures describe the inference that less specific utterances imply the negation of more specific utterances. For example, “Some of the students passed the test” is strictly less specific than “All of the students passed the test,” and therefore the use of the first utterance implicates that not all of the students passed. This is of course an example of a scalar implicature, in that there is a canonical scale, ordered according to logical strength, which both “some” and “all” fall on.

Not all specificity implicatures are naturally described as scalar implicatures. For example, consider the utterance “The object that I saw is green” in a context in which there are two green objects, one of which is a ball and one of which has an unusual and hard-to-describe shape. In this context, the utterance will be interpreted as describing the strangely shaped object, because the speaker could have said “The object that I saw is a ball” to uniquely pick out the ball (see Frank & Goodman, 2012 for experimental evidence for these implicatures). That is, in this context, there is an available utterance which is more specific than “green”, and as a result “green” receives a specificity implicature which is the negation of the more specific utterance. It is important to note that neither “green” nor “ball” is strictly logically stronger than the other; it is only in a particular context that one can be strictly more descriptive than the other. Thus, these utterances do not fall on a scale which is ordered according to logical strength.[2]

In general, specificity implicatures will arise in contexts in which there is a pair of utterances such that one utterance is more contextually specific than the other. To a first approximation, an utterance “A” is more contextually specific than “B” when the contextually-salient meanings consistent with “A” are a subset of those consistent with “B.” The use of the less specific utterance “B” will result in the inference that “A” is false. It is this more general phenomenon that the model will be explaining.

3.1 Derivation of specificity implicatures

This model can be used to derive specificity implicatures as follows. A rational speaker will use as specific of an utterance as possible in order to communicate with the literal listener; a more specific utterance is more likely to be interpreted correctly by the literal listener. If the speaker does not use a specific utterance, then this is evidence that such an utterance would not have communicated her intended meaning. The listener L_1 knows this, and (given the assumption of speaker knowledgeability) infers that the speaker must know that the more specific utterance is

[2] Though these utterances are logically incommensurable, it may still be possible to describe them as falling on an *ad-hoc* scale, as in Hirschberg (1985). While we will not be providing a direct argument against this analysis, our model obviates the need for a scalar representation in cases like this.

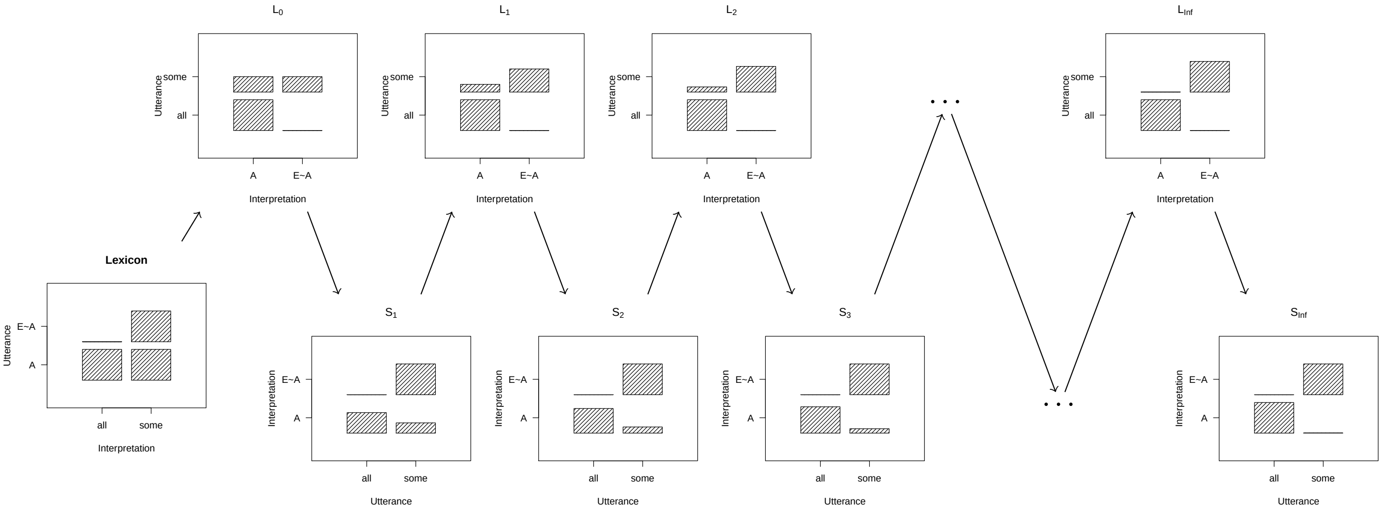


Figure 1: *Some* strengthening with $\mathrm{P}(\forall) = \frac{1}{2}$, $\mathrm{P}(\exists\neg\forall) = \frac{1}{2}$, $c(\text{all}) = c(\text{some}) = 0$, $\lambda = 1$

false. Therefore, a less specific utterance implies the negation of a more specific utterance for this listener.

To illustrate this reasoning, we will consider the simplest possible example in which specificity implicatures are possible. In this example, there are two utterances,

$$\mathcal{U} = \{\text{some, all}\},$$

and two meanings,

$$\mathcal{W} = \{\forall, \exists\neg\forall\},$$

where the intensions of the utterances are as usual:

$$[\![\text{some}]\!] = \{\forall, \exists\neg\forall\};$$
$$[\![\text{all}]\!] = \{\forall\}$$

Since it is common knowledge that the speaker knows the relevant world state, we can without loss of generality consider the observation and world variables to be equal, so that $o = w$, and drop w from the recursive equations (2)– (9). This allows the baseline model to be expressed as

$$L_0(o|u,\mathcal{L}) \propto \mathcal{L}(u,o)P(o), \tag{10}$$
$$U_n(u|o) = \log L_{n-1}(o|u) - c(u), \tag{11}$$
$$S_n(u|o) \propto e^{\lambda U_n(u|o)}, \tag{12}$$
$$L_n(o|u) \propto P(o)S_n(u|o), \tag{13}$$

for integers $n > 0$. For illustration, we take the prior on observations as uniform—$P(\exists\neg\forall) = P(\forall) = \frac{1}{2}$—the cost $c(u)$ of both utterances as identical (the specific value has no effect, and we treat it here as zero), and the softmax parameter $\lambda = 1$.[3]

[3] Changes in the prior on observations, utterance costs, and the softmax parameter change the precise values of the speaker and listener posteriors at various levels of recursion, but do not change the signature specificity-implicature pattern that the model exhibits.

Figure 1 depicts the listener and speaker posteriors $L_n(\cdot|u)$ and $S_n(\cdot|o)$ at increasing levels of recursion n for these parameter values. The lexicon matrix depicts the mapping of each possible utterance–world pair to a 0/1 value; each speaker (respectively listener) matrix should be read as a conditional distribution of utterances given interpretations (respectively interpretations given utterances), with bar height proportional to conditional probability (hence each row in each speaker or listener matrix sums to probability mass 1):

$$\begin{array}{lll} & \multicolumn{2}{c}{\textbf{Listener } n} \\ \hline \text{all} & L_n(\forall|\text{all}) & L_n(\exists\neg\forall|\text{all}) \\ \text{some} & L_n(\forall|\text{some}) & L_n(\exists\neg\forall|\text{some}) \\ & \forall & \exists\neg\forall \end{array} \qquad \begin{array}{lll} & \multicolumn{2}{c}{\textbf{Speaker } n} \\ \hline \forall & S_n(\text{all}|\forall) & S_n(\text{some}|\forall) \\ \exists\neg\forall & S_n(\text{all}|\exists\neg\forall) & S_n(\text{some}|\exists\neg\forall) \\ & \text{all} & \text{some} \end{array}$$

Crucially, while the literal listener interprets *some*, which rules out no worlds, entirely according to the prior (and hence as equiprobable as meaning $\forall$ and $\exists\neg\forall$), the speaker and listener both associate *some* increasingly strongly with $\exists\neg\forall$ as the pragmatic recursion depth increases.

One way to understand the fundamental reason for this behavior—the signature pattern of specificity implicature—is by considering the effect on one level of recursive inference on the listener's tendency to interpret *some* with unstrengthened meaning $\forall$. Let us denote $L_{n-1}(\forall|\text{some})$ by the probability p. Further, note that lexical constraints on the literal listener mean that $L_n(\exists\neg\forall|\text{all}) = 0$ always. This means that we can write, following Equations (11)–(13):

$$\begin{array}{lcc} \multicolumn{3}{l}{L_{n-1}} \\ \hline \text{all} & 1 & 0 \\ \text{some} & p & 1-p \\ & \forall & \exists\neg\forall \end{array} \quad \begin{array}{lcc} \multicolumn{3}{l}{U_n} \\ \hline \forall & 0 & \log p \\ \exists\neg\forall & -\infty & \log(1-p) \\ & \text{all} & \text{some} \end{array} \quad \begin{array}{lcc} \multicolumn{3}{l}{S_n} \\ \hline \forall & \frac{1}{1+p} & \frac{p}{1+p} \\ \exists\neg\forall & 0 & 1 \\ & \text{all} & \text{some} \end{array} \quad \begin{array}{lcc} \multicolumn{3}{l}{L_n} \\ \hline \text{all} & 1 & 0 \\ \text{some} & \frac{p}{2p+1} & \frac{1+p}{2p+1} \\ & \forall & \exists\neg\forall \end{array}$$

For all $p > 0$, the strict inequality $\frac{p}{2p+1} < p$ holds; therefore L_n is less inclined than L_{n-1} to interpret *some* as meaning $\forall$.

The above analysis assumed a uniform prior and $\lambda = 1$. The precise values of listener and speaker inferences are affected by these choices. A more exhaustive analysis of the behavior of this recursive reasoning system under a range of parameter settings is beyond the scope of the present paper, but the qualitative pattern of specificity implicature—that when pragmatic reasoning is formalized as recursive speaker–listener inference, more specific terms like *all* guide more general terms like *some* toward meanings not covered by the specific term—is highly general and robust to precise parameter settings. It is worth noting, however, that the value "greedy rationality" parameter λ affects the strength of the implicature when recursion depth is held constant. Figure 2 shows the tendency of the first pragmatic listener L_1 to interpret *some* as meaning $\forall$ (recall that for the literal listener, $L_0(\forall|\text{some}) = L_0(\exists\neg\forall|\text{some}) = \frac{1}{2}$ when the prior is uniform). This dependence on λ is due to L_1 modeling the first speaker S_1's degree of "greedy rationality". As greedy rationality increases, the strength of specificity implicature increases, to the extent that the possibility of $\forall$ interpretation for *some* can all but disappear after just one round of iteration with sufficiently high λ.

3.2 The symmetry problem

In addition to explaining specificity implicatures, the model provides a straightforward solution to the symmetry problem for scalar implicatures. As previously noted, on the standard account

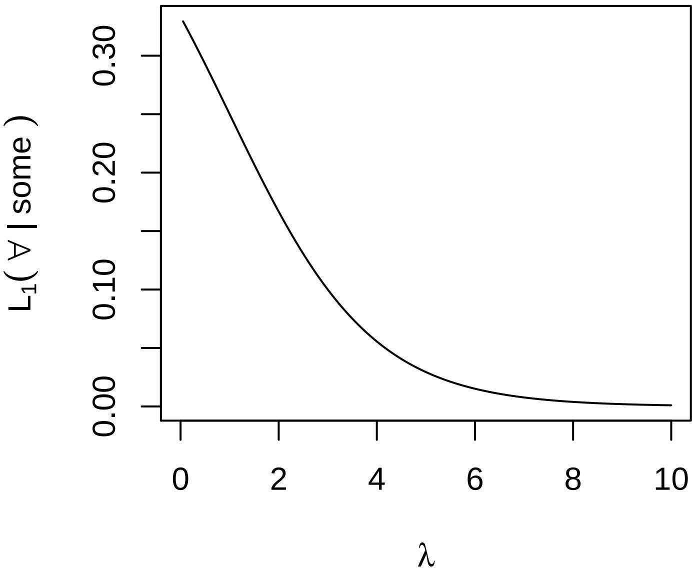


Figure 2: The degree of *some* strengthening as a function of the ”greedy rationality” parameter λ, with $P(\forall) = \frac{1}{2}$, $P(\exists\neg\forall) = \frac{1}{2}$, $c(\text{all}) = c(\text{some}) = 0$

of scalar implicatures, implicatures are computed with reference to a scale; lower utterances on the scale imply the negation of higher utterances on the scale. For example, the implicature for “some” is computed using the scale <“some”, “all”>, so that “some” implies the negation of “all.” The symmetry problem describes a problem with constructing the scales for the implicature computations: there are multiple consistent ways of constructing the scales, and different scales will give rise to different implicatures. The only formal requirement on a scale is that items higher on it be logically stronger than those lower on it. A possible scale for “some” is therefore <“some”, “some but not all”>. If this scale is used, “some” will imply that “some but not all” is not true, i.e. that “all” is true.

Fox & Katzir (2011) break the symmetry between “all” and “some but not all” by providing a theory of the alternative utterances which are considered during the computation of scalar implicatures. This theory posits that the set of scalar alternatives is computed via a set of combinatorial operations. That is, only the utterances which are constructed through these operations will be placed on the scale. The definition of these operations ensures that for each utterance on a scale, the set of utterances higher on the scale are consistent with each other. As a result, a consistent set of implicatures will be computed for each utterance.

The rational speech act model provides a different solution to this problem, which places weaker requirements on the set of alternative utterances. For the previous example, the model can include both “all” and “some but not all” as alternatives, and still derive the correct implicatures. It does so by assigning higher cost to “some but not all” than to “all.” Because “some but not all” is assigned a higher cost, it is less likely to be used to communicate *not all* than “all” is to communicate *all*. Thus, when the listener hears the utterance "some," they will reason that the speaker was likely to have intended to communicate *not all*: if the speaker had intended to communicate *all*, they would have used the utterance "all," but if they had intended to communicate *not all*, they would have been less likely to use "not all."

In general, this approach allows arbitrary sets of grammatical utterances to be considered as alternatives, without resulting in contradictory inferences, and while still preserving attested implicatures. The model will do this by assigning more complex utterances higher cost, and as a result weighing these more costly utterances less during pragmatic inference. Utterances that are more costly to the speaker are less likely to be used, because the speaker is rational. As an utterance becomes more and more costly, it becomes less and less salient to the speaker and listener as an alternative, and has less and less of an effect on the interpretation of other utterances.

4 Lexical uncertainty

4.1 M-implicatures

We will next consider a different type of pragmatic inference: M-implicatures. An M-implicature arises when there are two semantically equivalent utterances that differ in complexity. In general, the more complex utterance will receive a marked interpretation. The most straightforward way for an interpretation to be marked is for it to have low probability. Consider, for example, the following two sentences:

(i) John can finish the homework.

(ii) John has the ability to finish the homework.

These two sentences (plausibly) have the same literal semantic content, but they will typically not be interpreted identically. The latter sentence will usually be interpreted to mean that John will not finish the homework, while the former example does not have this implicature. Horn (1984) and Levinson (2000) cite a number of other linguistic examples which suggest that the assignment of marked interpretations to complex utterances is a pervasive phenomenon, in cases where there exist simpler, semantically equivalent alternatives.

Though M-implicatures describe a linguistic phenomenon, the reasoning that generates these implicatures applies equally to ad-hoc communication games with no linguistic component. Consider a one-shot speaker-listener signaling game with two utterances, SHORT and LONG (the costs of these utterances reflect their names), and two meanings, FREQ and RARE; nothing distinguishes the utterances other than their cost, and neither is assigned a meaning prior to the start of the game (so that effectively both have the all-*true* meaning). The speaker in this game needs to communicate one of the meanings; which meaning the speaker needs to communicate is sampled according to the prior distribution on these meanings (with the meaning FREQ having higher prior probability). The listener in turn needs to recover the speaker's intended meaning from their utterance. The speaker and listener will communicate most efficiently in this game if the speaker uses LONG in order to communicate the meaning RARE, and SHORT in order to communicate FREQ, and the listener interprets the speaker accordingly. That is, if the speaker and listener coordinate on this communication system, then the speaker will successfully transmit their intended meaning to the listener, and the expected cost to the speaker will be minimized. Bergen, Goodman & Levy (2012) find that in one-shot communication games of this sort, people do in fact communicate efficiently, suggesting that the pragmatic knowledge underlying M-implicatures is quite general and not limited to specific linguistic examples.[4]

4.1.1 Failure of rational speech acts model to derive M-implicatures

Perhaps surprisingly, our baseline rational speech-act model of Sections 2–3 is unable to account for speakers' and listeners' solution to the one-shot M-implicature problem. The behavior of the baseline model is shown in Figure 3; the model's qualitative failure is totally general across different settings of prior probabilities, utterance costs, and λ. The literal listener L_0 interprets both utterances identically, following the prior probabilities of the meanings. Crucially, L_0's interpretation distribution provides no information that speaker S_1 can leverage to associate either utterance with any specific meaning; the only thing distinguishing the utterances' expected utility is their cost. This leads to an across-the-board dispreference on the part of S_1 for LONG, but gives no starting point for more sophisticated listeners or speakers to break the symmetry between these utterances.

We will now formalize this argument; the following results will be useful in later discussions.

Lemma 1. *Let u, u' be utterances, and suppose $\mathcal{L}(u,w) = \mathcal{L}(u',w)$ for all worlds w. Then for all observations o and worlds w, $L_0(o,w|u,\mathcal{L}) = L_0(o,w|u',\mathcal{L})$.*

Proof. By equation 2,

$$
\begin{aligned}
L_0(o,w|u,\mathcal{L}) &= \frac{P(o,w)\mathcal{L}(u,w)}{\sum_{o',w'} P(o',w')\mathcal{L}(u,w')} && (14)\\
&= \frac{P(o,w)\mathcal{L}(u',w)}{\sum_{o',w'} P(o',w')\mathcal{L}(u',w')} && (15)\\
&= L_0(o,w|u',\mathcal{L}) && (16)
\end{aligned}
$$

where the equality in 15 follows from the fact that $\mathcal{L}(u,w) = \mathcal{L}(u',w)$ for all worlds w. □

Lemma 2. *Let u, u' be utterances, and suppose that $L_0(o,w|u,\mathcal{L}) = L_0(o,w|u',\mathcal{L})$ for all observations o and worlds w. Then for all observations o, worlds w, and $n \geq 0$, $L_n(o,w|u) = L_n(o,w|u')$.*

Proof. We will prove this by induction. Lemma 1 has already established the base case. Suppose that the statement is true up to $n-1 \geq 0$.

We will first consider the utility for speaker S_n. By equation 5,

$$
\begin{aligned}
U_n(u|o) - c(u') &= -D_{\mathrm{KL}}(P_o||L_{n-1}(\cdot|u)) - c(u) - c(u') && (17)\\
&= -D_{\mathrm{KL}}(P_o||L_{n-1}(\cdot|u')) - c(u') - c(u) && (18)\\
&= U_n(u'|o) - c(u) && (19)
\end{aligned}
$$

[4] The communication game considered in that paper differs slightly from the one considered here. In the experiments performed in that paper, there were three utterances available to the speaker, one of which was expensive, one of intermediate cost, and one cheap, and three possible meanings, one of which was most likely, one of intermediate probability, and one which was least likely. Participants in the experiment coordinated on the efficient mapping of utterances to meanings, i.e. the expensive utterance was mapped to the least likely meaning, and so on.

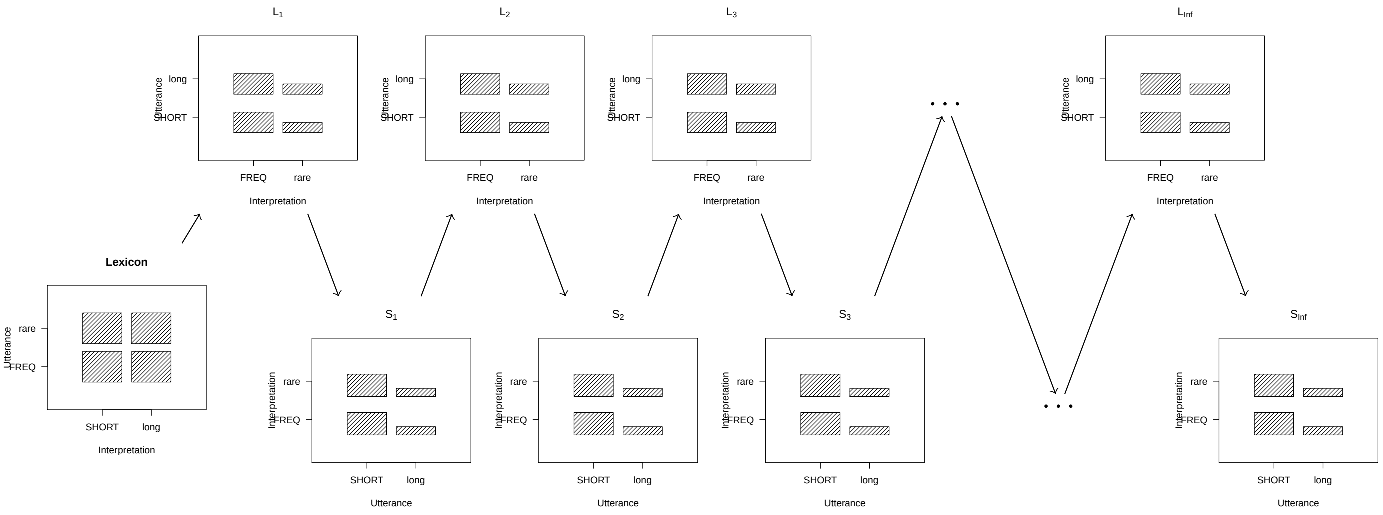


Figure 3: The failure of the basic model to derive M-implicature (illustrated here for $P(\text{FREQ}) = \frac{2}{3}$, $P(\text{rare}) = \frac{1}{3}$, $\lambda = 3$, $c(\text{SHORT}) = 1$, $c(\text{long}) = 2$)

It follows from equation 8 that:

$$S_n(u|o) = \frac{e^{\lambda U_n(u|o)}}{\sum_{u_i} e^{\lambda U_n(u_i|o)}} \tag{20}$$

$$= \frac{e^{\lambda(U_n(u'|o) - c(u) + c(u'))}}{\sum_{u_i} e^{\lambda U_n(u_i|o)}} \tag{21}$$

$$= S_n(u'|o) \cdot e^{\lambda(c(u') - c(u))} \tag{22}$$

In other words, for all observations o, $S_n(u|o)$ and $S_n(u'|o)$ differ by a constant factor determined by the difference of the utterances' costs.

We will now show the equivalence of listeners $L_n(\cdot|u)$ and $L_n(\cdot|u')$. By equation 9,

$$L_n(o,w|u) = \frac{P(o,w)S_n(u|o)}{\sum_{o',w'} P(o',w')S_n(u|o')} \tag{23}$$

$$= \frac{P(o,w)S_n(u'|o)e^{\lambda(c(u') - c(u))}}{\sum_{o',w'} P(o',w')S_n(u'|o')e^{\lambda(c(u') - c(u))}} \tag{24}$$

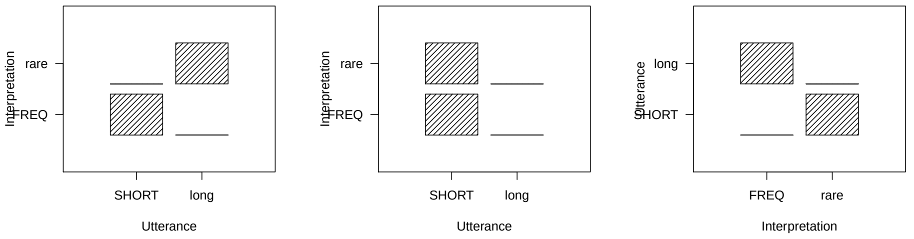
$$= L_n(o,w|u') \tag{25}$$

□

Together, these lemmas show that if two utterances have the same literal meanings, then they will be interpreted identically at all levels of the speaker-hearer recursion in the rational speech acts model.

4.2 The multiple equilibrium problem

Our baseline model's failure for M-implicature is in fact closely related to a more general problem from game theory, the multiple equilibrium problem for signalling games (Cho & Kreps, 1987;



(a) Separating Pareto-optimal

(b) Pooling 1

(c) Separating Pareto-suboptimal

Figure 4: Multiple equilibria (speaker matrices) for the M-implicature signaling game

Rabin, 1990). In a typical signalling game, a subset of the agents in the game each receive a type, where this type is revealed only to the agent receiving it; in the settings being considered in this paper, each speaker has a type, which is the meaning that they want to communicate. The goal of the listener is to correctly guess the type of the speaker based on the signal that they send.

To describe the multiple equilibrium problem for such games, we first need to introduce the relevant notion of equilibrium. Loosely speaking, the equilibria for a game describe the self-consistent ways that the game can be played. The simplest equilibrium concept in game theory is the Nash equilibrium (Fudenberg & Tirole, 1991; Myerson, 2013; Nash et al., 1950). For games with two agents A and B, a pair of strategies (σ_A, σ_B), which describe how each agent will play the game, are a Nash equilibrium if neither agent would benefit by unilaterally changing their strategy; that is, the strategies are an equilibrium if, fixing σ_B, there is no strategy for A that would improve the outcome of the game for A, and vice-versa.

The relevant notion of equilibrium for signalling games is the Bayesian Nash equilibrium (Harsanyi, 1967), which in addition to the requirements imposed by the definition of the Nash equilibrium also imposes consistency constraints on the beliefs of the agents. In particular, given the prior distribution over types, and the agents' strategies (which define the likelihood of taking actions given a player type), the agents must use Bayes' rule to compute their posterior distribution over types after observing an action. Each agent's strategy must also be rational given their beliefs at the time that they take the action, in the sense that the strategy must maximize their expected utility. The multiple equilibrium problem arises in a signalling game when the game has multiple Bayesian Nash equilibria. This occurs when the agents can devise multiple self-consistent communication systems given the constraints of the game. That is, given the assumption that the other agents are using the communication system, it will not be rational for one agent to unilaterally start using a different communication system.

The multiple equilibrium problem can be illustrated concretely using the game above. This game has two general classes of equilibria, illustrated in Figure 4. In the first class, which are called the *separating equilibria*, successful communication occurs between the speaker and listener, but their communication system may be suboptimal from an information-theoretic perspective. In the first such equilibrium, the speaker chooses LONG when they want to communicate RARE, and SHORT when they want to communicate FREQ (Figure 4a). Given these strategies, the listener knows how to interpret each utterance: LONG will be interpreted as RARE— conditional on hearing

LONG, the only possibility is that it was produced by the agent wanting to communicate RARE—and similarly SHORT will be interpreted as FREQ. This is clearly an equilibrium, because neither speaker will successfully communicate their intended meaning if they unilaterally change their strategy; for example, if the speaker wanting to communicate RARE switches to using SHORT, then they will be interpreted as intending FREQ. A second separating equilibrium is also possible in this game. Under this equilibrium, the speaker-utterance pairs are reversed, so that the agent intending to communicate RARE uses SHORT, and the agent intending FREQ uses LONG (Figure 4c). This is inefficient — in expectation, it will be more expensive than the previous equilibrium for the speaker — but it is nonetheless an equilibrium, because neither speaker can unilaterally change strategies without failing to communicate.

The second type of equilibrium in this game, known as a *pooling equilibrium*, is still more deficient than the inefficient separating equilibrium, and it is the one that is most closely related to the problem for our initial model of pragmatic inference. In one pooling equilibrium, the speaker chooses the utterance SHORT, independent of the meaning that they want to communicate (Figure 4b). Because the speakers always choose SHORT, this utterance communicates no information about the speaker's intended meaning, and the listener interprets this utterance according to the prior distribution on meanings. Assuming that the utterance LONG is also interpreted according to the prior, it will never be rational for the speaker to choose this utterance.[5] Thus this is indeed an equilibrium.

These arguments demonstrate that under the standard game-theoretic signalling model, speakers and listeners are not guaranteed to arrive at the efficient communication equilibrium. Rather, there is the possibility that they will successfully communicate but do so inefficiently, with cheaper utterances interpreted as referring to less likely meanings. There is also the possibility that they will fail to communicate at all, in the case that all speakers choose the cheapest available utterance. However, M-implicatures demonstrate that at least in certain cases, people are able to systematically coordinate on the efficient strategies for communication, even when semantics provides no guide for breaking the symmetries between utterances. Thus, there is something to account for in people's strategic and pragmatic reasoning beyond what is represented in standard game-theoretic models or in our initial model of pragmatic reasoning.

In recent work in linguistics, there have generally been three approaches to accounting for these reasoning abilities. The first approach uses the notion of a *focal point* for equilibria (Parikh, 2000). On this approach, people select the efficient equilibrium in signalling games because it is especially salient; the fact that it is salient makes each agent expect other agents to play it, which in turn makes each agent more likely to play it themselves. While this approach does derive the efficient equilibrium for communication games, it is not entirely satisfactory, since it does not provide an independent account of salience in these games — precisely the feature which allows the agents to efficiently communicate under this approach.

An alternative approach has been to derive the efficient equilibrium using evolutionary game theory, as in De Jaegher (2008); Van Rooy (2004). These models show that given an appropriate evolutionary dynamics, inefficient communication systems will evolve towards more efficient systems among collections of agents. While these models may demonstrate how efficient semantic

[5] Note that because in this equilibrium the speaker never uses one of the two utterances, the listener cannot interpret the never-used utterance by Bayesian conditioning, because it is not possible to condition on a probability 0 event. As a result, standard game-theoretic models need to separately specify the interpretation of probability 0 signals. We will return to this issue below.

conventions can evolve among agents, they do not demonstrate how agents can efficiently communicate in one-shot games. Indeed, in the relevant setting for M-implicatures, the agents begin with an inefficient communication system — one in which the semantics of their utterances does not distinguish between the meanings of interest — and must successfully communicate within a single round of play. There is no room for selection pressures to apply in this setting.

Finally, Franke (2009), Jäger (2012), and Franke & Jäger (2013) have derived M-implicatures in the Iterated Best Response (IBR) and Iterated Quantal Response (IQR) models of communication, which are closely related to the rational speech act model considered in the previous section. The naive versions of these models do not derive M-implicatures, for reasons that are nearly identical to why the rational speech act model fails to derive them. In the IBR model, players choose strategies in a perfectly optimal manner. Because the expensive utterance in the Horn game is strictly worse than the cheap utterance — it is more expensive and has identical semantic content – an optimal speaker will never use it. As a result, in the naive IBR model, the speaker chooses the expensive utterance with probability 0, and no coherent inference can be drawn by the listener if they hear this utterance; interpreting this utterance would require them to condition on a probability 0 event. Franke (2009) and Jäger (2012) show how to eliminate this problem in the IBR model and correctly derive M-implicatures. They propose a constraint on how listeners interpret probability 0 utterances, and show that this constraint results in the efficient equilibrium. This proposal cannot be extended to the rational speech acts model, because it relies on the expensive utterance being used with probability 0; in the rational speech acts model, agents are only approximately rational, and as a result, every utterance is used with positive probability.

As in the rational speech acts model, agents are only approximately rational in the IQR model, and the IBR derivation of M-implicatures similarly does not extend to this model. Franke & Jäger (2013) therefore provide an alternative extension of the IQR model which derives M-implicatures. Under this proposal, agents who receive low utility from all of their available actions engage in more exploratory behavior. In a Horn game, the speaker who wants to communicate the meaning RARE starts out with a low expected utility from all of their actions: no matter which utterance they choose, the listener is unlikely to interpret them correctly. As a result, this speaker will engage in more exploratory behavior — i.e., behave less optimally with respect to their communicative goal — and will be more likely to choose the suboptimal expensive utterance. This is sufficient to break the symmetry between the cheap and expensive utterances, and derive the M-implicature.

Unlike the proposed modification of the IBR model, Franke & Jäger (2013)'s proposed derivation of M-implicatures within the IQR model would extend straightforwardly to the rational speech acts model. We will nonetheless be proposing an alternative extension to the rational speech acts model. This is for several reasons. First, the derivation within the IQR model depends on the empirical assumption that agents with worse alternatives available to them will choose among these alternatives less optimally than agents with better alternatives available. Though this is a reasonable assumption, it may turn out to be empirically false; to our knowledge, it has not been experimentally evaluated. As a general claim about how agents make decisions, it will have consequences for other areas of psychological theorizing as well. Second, the derivation of M-implicatures which we present can be extended to explain a number of other phenomena, which will be discussed in later sections. These explanations will hinge on features which are distinctive to our proposed extension of the rational speech acts model.

4.3 The lexical-uncertainty model

In the previous version of the model, it was assumed that the lexicon $\mathcal{L}$ used by the speaker and listener was fixed. For every utterance u, there was a single lexical entry $\mathcal{L}(u,\cdot)$ that gave the truth function for u. This fixed lexicon determined how the literal listener would interpret each utterance.

In the current version of the model, we introduce *lexical uncertainty*, so that the fixed lexicon is replaced by a set of lexica Λ over which a there is a probability distribution $P(\mathcal{L})$. This distribution represents sophisticated listeners' and speakers' uncertainty about how the literal listener will interpret utterances. (Alternative formulations of lexical uncertainty may be clear to the reader; in Appendix B we describe two and explain why they don't give rise do the desired pragmatic effects.)

Introducing lexical uncertainty generalizes the previous model; the base listener L_0 remains unchanged from equation 2, i.e. this listener is defined by:

$$L_0(o,w|u,\mathcal{L}) \propto \mathcal{L}(u,w)P(o,w) \tag{26}$$

for every lexicon $\mathcal{L} \in \Lambda$. The more sophisticated speakers and listeners, S_n and L_n for $n \geq 1$, are defined by:

$$U_1(u|o,\mathcal{L}) = -D_{KL}(P_o||L_0^{u,\mathcal{L}}) - c(u), \tag{27}$$

where $L_k^{u,\mathcal{L}}$ is the level-k listener's posterior distribution on o and w conditional on utterance u and lexicon $\mathcal{L}$,

$$\begin{aligned}
S_1(u|o,\mathcal{L}) &\propto e^{\lambda U_1(u|o,\mathcal{L})}, && \quad (28)\\
L_1(o,w|u) &\propto P(o,w) \sum_{\mathcal{L}\in\Lambda} P(\mathcal{L})S_1(u|o,\mathcal{L}), && \quad (29)\\
U_n(u|o) &= -D_{KL}(P_o||L_{n-1}^{u}) - c(u) \qquad \text{for } n>1, && \quad (30)\\
S_n(u|o) &\propto e^{\lambda U_n(u|o)} \qquad \text{for } n>1, && \quad (31)\\
L_n(o,w|u) &\propto P(o,w)S_n(u|o) \qquad \text{for } n>1.^{6} && \quad (32)
\end{aligned}$$

In order for the normalization of Equation 26 and the KL divergence of Equation 27 to be well-defined we must place two restrictions on each $\mathcal{L} \in \Lambda$.

[6] It is possible to define the lexical-uncertainty model more concisely by replacing Equations (27)–(32) with the following three equations:

$$\begin{aligned}
U_n(u|o,w,\mathcal{L}) &= -D_{KL}(P_o||L_{n-1}^{u,\mathcal{L}}) - c(u). && \quad \text{(i)}\\
S_n(u|o,w,\mathcal{L}) &\propto e^{\lambda U_n(u|o,w,\mathcal{L})}, && \quad \text{(ii)}\\
L_n(o,w|u,\mathcal{L}) &\propto \sum_{\mathcal{L}'\in\Lambda} P(o,w)P(\mathcal{L}')S_n(u|o,w,\mathcal{L}'), && \quad \text{(iii)}
\end{aligned}$$

Once the first marginalization over lexica occurs at the L_1 level, higher-level speaker and listener distributions lose their dependence on the lexicon $\mathcal{L}$ being conditioned on, since there is no dependence on $\mathcal{L}$ in the right-hand side of equation (iii). In this paper we rely on the less concise definitions provided in the main text, however, on the belief that they are easier to follow than those in Equations (i)–(iii).

1. Each utterance must receive a non-contradictory interpretation. Formally, for each utterance u and each lexicon $\mathcal{L} \in \Lambda$ there must exist a world w such that $\mathcal{L}(u,w) > 0$.

2. For any observation there is an utterance which includes the speaker's belief state in its support. Formally, for each observation o and each lexicon $\mathcal{L} \in \Lambda$ there exists (at least) one utterance u such that $\mathcal{L}(u,w) > 0$ for any w with $P(w|o) > 0$.

Satisfying the first of these restrictions is straightforward. We have considered four approaches to constructing Λ that satisfy the second restriction, each of which result in qualitatively similar predictions for all of the models considered in this paper. In the first of these approaches, the global constraint of restriction 2 is simply imposed on each lexicon by fiat; any lexicon which does not satisfy this condition is assigned probability 0. In the second of these approaches, the truth-conditional semantics of each utterance is slightly weakened. When an utterance u is false at a world state w, we define $\mathcal{L}(u,w) = 10^{-6}$ (or any smaller, positive number). In this case, each utterance always assigns at least a small amount of mass to each world state, immediately satisfying restriction 2. In the third approach we assume that there is some, much more complex, utterance that could fully specify any possible belief state. That is, for any o there is an utterance u_o such that $\mathcal{L}(u_o,\cdot)$ coincides with the support of $P(w|o)$ in every lexicon $\mathcal{L} \in \Lambda$. The utterances u_o may be arbitrarily expensive, so that the speaker is arbitrarily unlikely to use them; they still serve to make the KL divergence well-defined. This approach captures the intuition that real language is infinitely expressive in the sense that any intended meaning can be conveyed by some arbitrarily complex utterance. The fourth approach is a simplification of the previous one: we collapse the u_o into a single utterance u_{null} such that $\mathcal{L}(u_{null},w) = 1$ for every world w. Again u_{null} is assumed to be the most expensive utterance available. In the remainder we adopt this last option as the clearest for presentational purposes. In the models we consider in the remainder of this paper, u_{null} never becomes a preferred speaker choice due to its high cost, though it is possible that for other problems u_{null} may turn out to be an effective communicative act. We leave the question of whether this is a desirable feature of our model for future work.

The above restrictions leave a great deal of flexibility for determining Λ; in practice we adopt the most complete Λ that is compatible with the base semantics of our language. If we begin with a base SEMANTIC LEXICON, $\mathcal{L}_S$, for the language (i.e. the lexicon that maps each utterance to its truth function under the language's semantics) we can define Λ by a canonical procedure of sentential enrichment: Call the utterance meaning $\mathcal{L}(u,\cdot)$ a *valid refinement* of $\mathcal{L}_S$ if: $\forall w\ \mathcal{L}_S(u,w) = 0 \implies \mathcal{L}(u,w) = 0$, and, $\exists w\ \mathcal{L}(u,w) > 0$. Define $\tilde{\Lambda}$ to consist of all lexica $\mathcal{L}$ such that each utterance meaning is a valid refinement of the meaning in $\mathcal{L}_S$; define the ENRICHMENT Λ of $\mathcal{L}_S$ to be $\tilde{\Lambda}$ with an additional utterance u_{null} added to each lexicon, such that $\mathcal{L}(u_{null},w) = 1$ for every world w.

4.4 Specificity implicature under lexical uncertainty

Before demonstrating how the lexical-uncertainty model derives M-implicature (which we do in Section 4.5), in this section we walk the reader through the operation of the lexical-uncertainty model for a simpler problem: the original problem of specificity implicature, which the revised lexical-uncertainty model also solves. The setup of the problem remains the same, with (equal-cost) utterance set $\mathcal{U} = \{\text{some}, \text{all}\}$, meanings $\mathcal{W} = \{\forall, \exists\neg\forall\}$, and literal utterance meanings—semantic lexicon $\mathcal{L}_S$ in the terminology of Section 4.3—$[\![\text{some}]\!] = \{\forall, \exists\neg\forall\}, [\![\text{all}]\!] = \{\forall\}$. The enrichment procedure gives Λ consisting of:

$$\mathcal{L}_1 = \left\{ \begin{array}{ll} [\![\text{all}]\!] & = \{\forall\} \\ [\![\text{some}]\!] & = \{\exists\neg\forall, \forall\} \\ [\![u_{null}]\!] & = \{\exists\neg\forall, \forall\} \end{array} \right\} \quad \mathcal{L}_2 = \left\{ \begin{array}{ll} [\![\text{all}]\!] & = \{\forall\} \\ [\![\text{some}]\!] & = \{\exists\neg\forall\} \\ [\![u_{null}]\!] & = \{\exists\neg\forall, \forall\} \end{array} \right\} \quad \mathcal{L}_3 = \left\{ \begin{array}{ll} [\![\text{all}]\!] & = \{\forall\} \\ [\![\text{some}]\!] & = \{\forall\} \\ [\![u_{null}]\!] & = \{\exists\neg\forall, \forall\} \end{array} \right\}$$

and we make the minimal assumption of a uniform distribution over Λ: $P(\mathcal{L}_1) = P(\mathcal{L}_2) = P(\mathcal{L}_3) = \frac{1}{3}$. Note that *some* can be enriched to either $\exists\neg\forall$ or to $\forall$, and before pragmatic inference gets involved there is no preference among either those two or an unenriched meaning.

We can now compute the behavior of the model. Since it is common knowledge that the speaker knows the relevant world state, we can once again let $o = w$ and drop w from the recursive equations, so that the lexical-uncertainty model of Equations (26)–(32) can be expressed as

$$\begin{aligned} L_0(o|u,\mathcal{L}) &\propto \mathcal{L}(u,o)P(o), && (33) \\ U_1(u|o,\mathcal{L}) &= \log L_0(o|u,\mathcal{L}) - c(u), && (34) \\ S_1(u|o,\mathcal{L}) &\propto e^{\lambda U_1(u|o,\mathcal{L})}, && (35) \\ L_1(o|u) &\propto P(o) \sum_{\mathcal{L}\in\Lambda} P(\mathcal{L}) S_1(u|o,\mathcal{L}), && (36) \\ U_n(u|o) &= \log L_{n-1}(o|u) - c(u) \qquad \text{for } n > 1, && (37) \\ S_n(u|o) &\propto e^{\lambda U_n(u|o)} \qquad \text{for } n > 1, && (38) \\ L_n(o|u) &\propto P(o)S_n(u|) \qquad \text{for } n > 1. && (39) \end{aligned}$$

Figure 5 shows the listener and speaker posterior distributions at varying levels of recursion. At the L_0 literal-listener and S_1 first-speaker levels, different inferences are drawn conditional on the lexicon entertained: the three lexica $\mathcal{L}_1$ through $\mathcal{L}_3$ are stacked top to bottom in the leftmost panel, and the dependencies among lexicon-specific inferences are indicated with arrows between panels. Up through S_1, each lexicon-specific recursive inference chain operates indistinguishably from that of the baseline model, except that an enriched lexicon rather than the base semantic lexicon of the language is used throughout.

The specificity implicature first appears at the level of the listener L_1, who is reasoning about the speaker S_1. The listener computes their posterior distribution over the speaker's intended meaning by marginalizing over the possible lexica that the speaker may have been using (Equation (36)). As can be seen in the second column of the third panel of Figure 5, S_1's posterior supports three different possible interpretations of *some*. Under the lexicon in which *some* has been enriched to mean $\forall$ (bottom subpanel), *some* should be interpreted to categorically mean $\forall$; under the lexicon in which *some* has been enriched to mean $\exists\neg\forall$ (middle subpanel), *some* should be interpreted to categorically mean $\exists\neg\forall$. Under the lexicon in which *some* remains unenriched, *some* should be preferentially interpreted as $\exists\neg\forall$ due to blocking of $\forall$ by *all*, exactly as in the baseline model. Thus in the final mixture of lexica determining the overall interpretive preferences of L_1, there is an overall preference of *some* to be interpreted as $\exists\neg\forall$; this preference can get further strengthened through additional speaker–listener iterations, exactly as in the baseline model. Thus specificity implicatures still go through under lexical uncertainty.

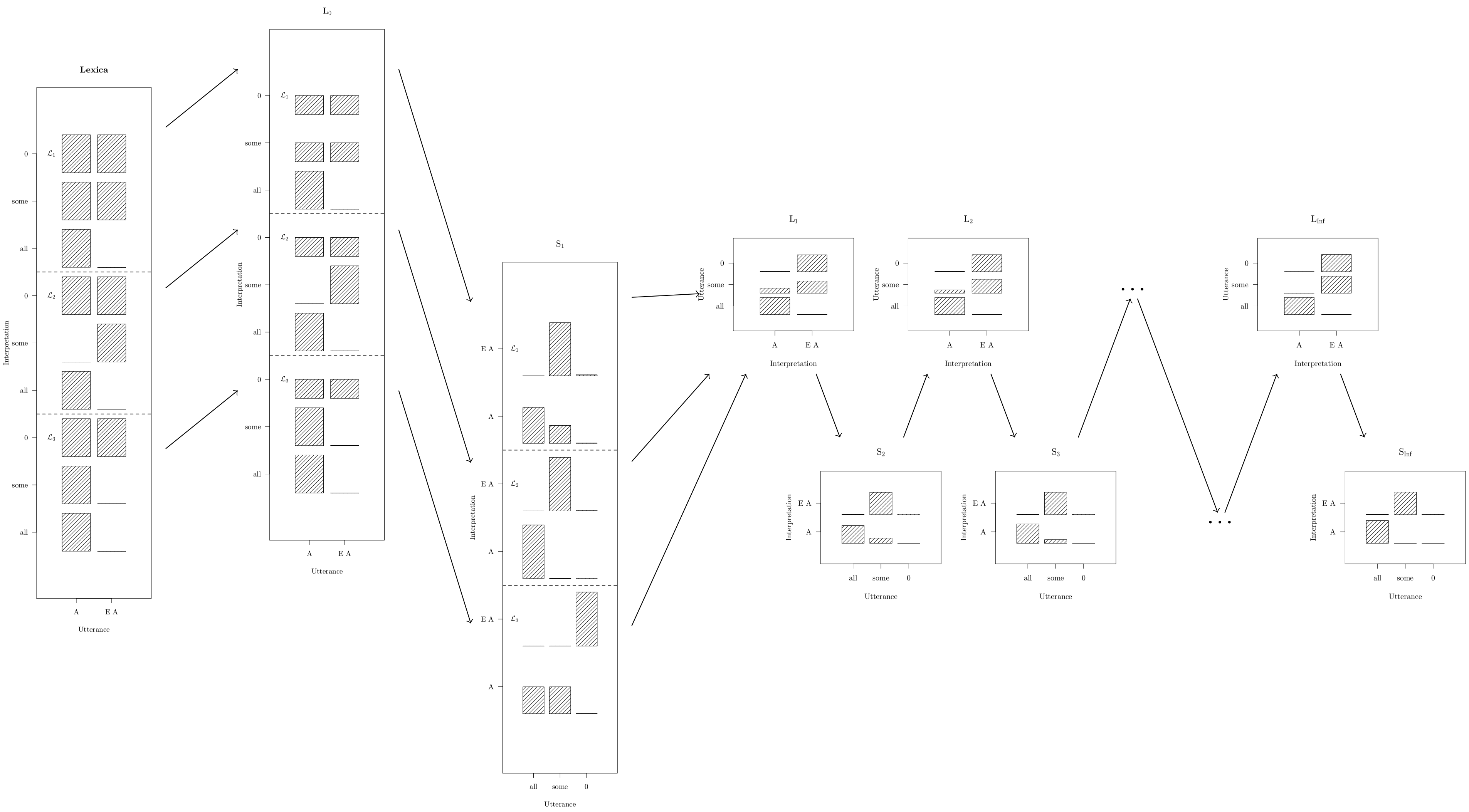


Figure 5: Specificity implicatures under lexical uncertainty, shown here with $\mathrm{P}(\forall) = \frac{1}{2}$, $\mathrm{P}(\exists\neg\forall) = \frac{1}{2}$, $c(\text{all}) = c(\text{some}) = 1$, $c(\emptyset) = 5$, $\lambda = 1$

4.5 Derivation of M-implicature under lexical uncertainty

We now show how lexical uncertainty allows the derivation of one-shot M-implicatures. We consider the simplest possible M-implicature problem of two possible meanings to be communicated—one higher in prior probability (FREQ) than the other (RARE)—that could potentially be signaled by two utterances—one less costly (SHORT) than the other (LONG). The semantic lexicon of the language is completely unconstrained:

$$\mathcal{L}_S = \left\{ \begin{array}{ll} [\![\text{SHORT}]\!] & = \{\text{FREQ}, \text{RARE}\} \\ [\![\text{LONG}]\!] & = \{\text{FREQ}, \text{RARE}\} \end{array} \right\}$$

Each utterance has three possible enrichments—$\{\text{FREQ}, \text{RARE}\}$, $\{\text{FREQ}\}$, and $\{\text{RARE}\}$—leading to nine logically possible enriched lexica. We make the minimal assumption of taking Λ to be this complete set of nine, illustrated in the first panel of Figure 6, and putting a uniform distribution over Λ.

Because utterance costs play no role in the literal listener's inferences, L_0 is completely symmetric in the behavior of the two utterances (second panel of Figure 6). However, the variety in lexica gives speaker S_1 resources with which to plan utterance use efficiently. The key lexica in question are the four in which the meaning of only one of the two utterances is enriched: $\mathcal{L}_2$, $\mathcal{L}_3$, $\mathcal{L}_4$, and $\mathcal{L}_7$. $\mathcal{L}_2$ and $\mathcal{L}_7$ offer the speaker the partial associations LONG–RARE and SHORT–FREQ, respectively, whereas $\mathcal{L}_3$ and $\mathcal{L}_4$ offer the opposite: LONG–FREQ and SHORT–RARE, respectively. Crucially, the former pair of associations allows greater expected speaker utility, and thus undergo a stronger specificity implicature in S_1, than the latter pair of associations.

This can be seen most clearly in the contrast between $\mathcal{L}_2$ and $\mathcal{L}_3$. The speaker S_1 forms a stronger association of LONG to RARE in $\mathcal{L}_2$ than of LONG to FREQ in $\mathcal{L}_3$. This asymmetry arises because the value of precision varies with communicative intention. A speaker using $\mathcal{L}_2$ can communicate RARE precisely by using LONG, and will not be able to effectively communicate this meaning by using the vague utterance SHORT. Thus, this speaker will be relatively likely to use LONG to communicate RARE. In contrast, LONG will communicate FREQ precisely under $\mathcal{L}_3$, but this meaning can also be communicated effectively with the utterance SHORT. Thus, the speaker using $\mathcal{L}_3$ will be less likely to choose LONG.

When the first pragmatic listener L_1 takes into account the variety of S_1 behavior across possible lexica (through the marginalization in Equation (36)), the result is a weak but crucial LONG–RARE association. Further levels of listener–speaker recursion amplify this association toward increasing categoricality. (The parameter settings in Figure 6 are chosen to make the association at the L_1 level relatively visible, but the same qualitative behavior is robust for all finite $\lambda > 1$.) Simply by introducing consideration of multiple possible enrichments of the literal semantic lexicon of the language, lexical uncertainty allows listeners and speakers to converge toward the M-implicature equilibrium that is seen not only pervasively in natural language but also in one-shot rounds of simple signaling games.

4.6 Ignorance as a marked state

The lexical-uncertainty model introduced earlier in this section provided a novel means by which speakers and listeners in one-shot communication games align forms and meanings in terms of

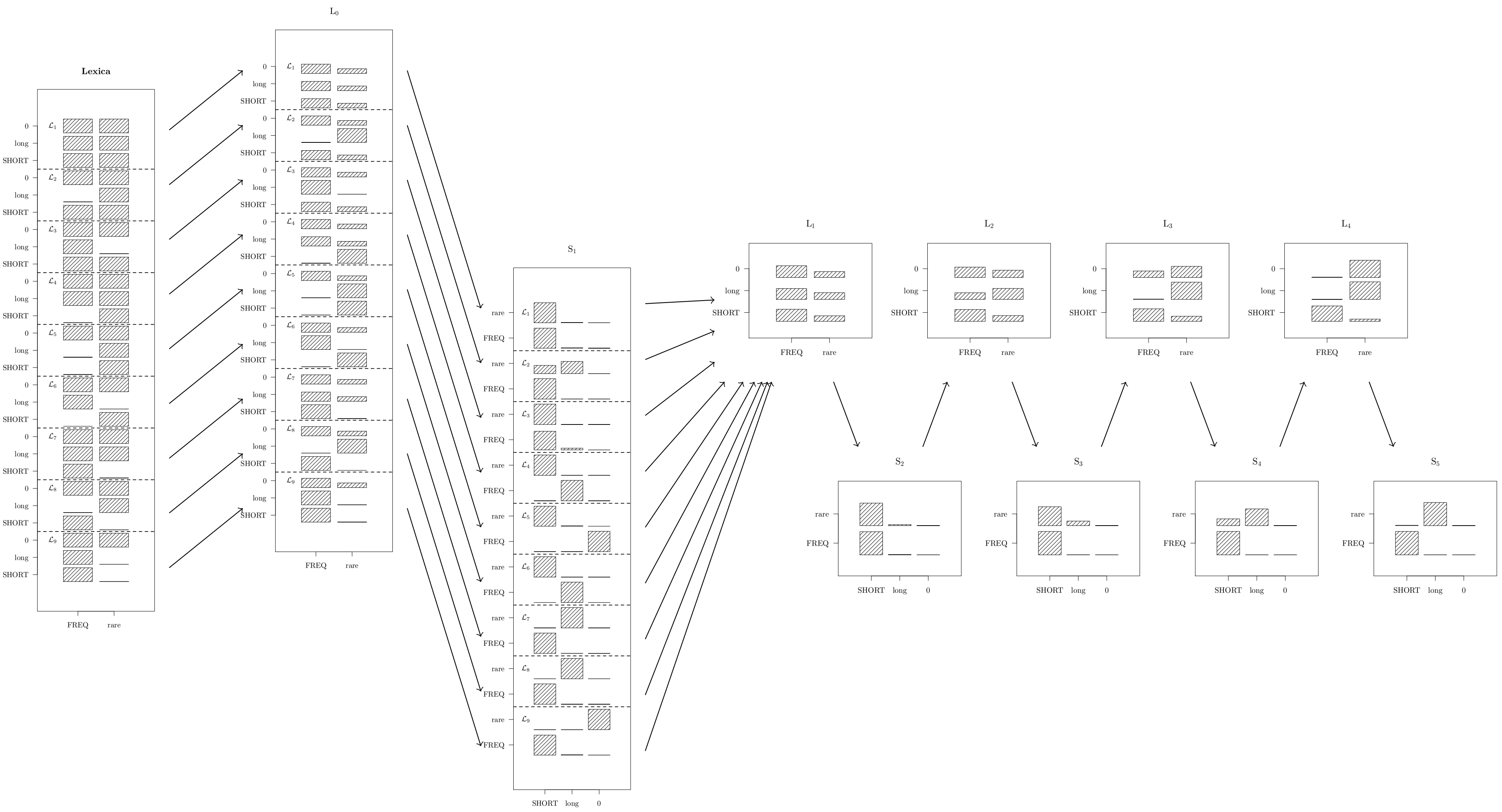


Figure 6: Deriving M-implicature with $P(\text{FREQ}) = \frac{2}{3}$, $P(\text{rare}) = \frac{1}{3}$, $\lambda = 4$, $c(\text{SHORT}) = 1$, $c(\text{long}) = 2$, $c(\emptyset) = 5$.

what can be thought of as two different types of *markedness*: cost of forms and prior probabilities, or frequencies, of meanings. Perhaps remarkably, a third type of markedness emerges as a side effect of this model that can explain a particularly vexing class instance of implicature, most famously exemplified by the sentence pair below:

(i) Some or all of the students passed the test.

(ii) Some of the students passed the test.

As discussed in Section 3, (ii) has a specificity implicature that strengthens the literal meaning of "some" to an understood meaning of "some but not all". The implicatures of (i) differ crucially in two ways. First, as noted by Gazdar (1979, see also Chierchia et al., 2012), (i) lacks the basic specificity implicature of (ii). Second, (i) seems to possess an *ignorance* implicature: namely, that the speaker is not sure whether or not all the students passed the test.

Accounting for why the specificity implicature is lacking and how the ignorance implicature comes about has become a problem of considerable prominence in recent semantic and pragmatic theory (Fox, 2007,1; Meyer, 2013; Russell, 2012). This is for several reasons. First, the sentence in (i) violates Hurford's constraint (Hurford, 1974), according to which a disjunction is infelicitous if one of its disjuncts entails the other. In this case, because "all" entails "some," the constraint incorrectly predicts that the sentence should be infelicitous. For closely related reasons, neo-Gricean theories — as well as the rational speech acts model from Section 2 — cannot derive the implicatures associated with this sentence. A disjunction which violates Hurford's constraint will be semantically equivalent to one of its disjuncts (i.e. the weaker one); in this case, the expression "some or all" is semantically equivalent to "some." As previously discussed, the rational speech acts model, and neo-Gricean models more generally, cannot derive distinct pragmatic interpretations for semantically equivalent expressions.

4.6.1 An empirical test of ignorance implicature

Before proceeding further, a note regarding the available data is called for. To the best of our knowledge, the only data adduced in the literature in support of the claim that sentences like (i) possess ignorance implicatures have been introspective judgments by the authors of research articles on the phenomenon in question. It is therefore worth briefly exploring exactly how this claim might be more objectively tested and thus verified or disconfirmed. In our view, the claim that "some or all" sentences such as (i) possess an ignorance implicature that corresponding sentences such as (ii) do not should make the following empirically testable prediction: that of sentence pairs like ((i)–(ii)) differing only in TARGET QUANTIFIER "some or all" versus "some", comprehenders should be less likely to conclude that the speaker knows, and/or conclude that the speaker is less likely to know, that substitution of the target quantifier with both (a) "all", and (b) "not all", for the "some or all" variant than for the "some" variant. To test this prediction, we ran a brief experiment that involved presenting speakers with paragraphs of the following type, each in one of two variants:

> Letters to Laura's company almost always have checks inside. Today Laura received 10 letters. She may or may not have had time to check all of the letters to see if they have checks. You call Laura and ask her how many of the letters have checks inside. She says, "{Some/Some or all} of the letters have checks inside."

Participants were asked two questions:

- *How many letters did Laura look inside?* Answers to this question confirmed (a) above: significantly fewer participants answered *10* in the “some” condition than in the “some or all” condition.
- *Of the letters that Laura looked inside, how many had checks in them?* Answers to this question confirmed (b) above: significantly fewer participants gave the same number as an answer to both this and the preceding question in the “some” condition than in the “some or all “condition.

We are now on more solid ground in asserting that “some or all” triggers an ignorance implicature that is lacked by “some” and that needs to be explained, and proceed to derive this ignorance implicature within our lexical-uncertainty model. (Further details of this experiment can be found in Appendix A.)

4.6.2 Deriving ignorance implicatures

To show how our model derives ignorance implicature for the “some or all” case, we first lay out assumptions about the set of world and observation states, the prior over these states, the contents of the semantic lexicon, and utterance costs:

		w	
	$P(o,w)$	$\forall$	$\exists\neg\forall$
	$\forall$	$\frac{1}{3}$	0
o	$?$	$\frac{1}{6}$	$\frac{1}{6}$
	$\exists\neg\forall$	0	$\frac{1}{3}$

$$\mathcal{L}_S = \left\{ \begin{array}{ll} [\![\text{all}]\!] & = \{\forall\} \\ [\![\text{some}]\!] & = \{\exists\neg\forall, \forall\} \\ [\![\text{some or all}]\!] & = \{\exists\neg\forall, \forall\} \end{array} \right\}$$

u	$c(u)$
all	0
some	0
some or all	1

Exactly as before in our treatment of specificity implicature in Sections 3 and 4.4, we assume two possible world states: $\mathcal{W} = \{\forall, \exists\neg\forall\}$. In order to capture the notion of possible speaker ignorance, however, we have relaxed the assumption of a one-to-one mapping between speaker observation state and world state, and allow three observation states: $\exists\neg\forall$, $\forall$, and a third, “ignorance” observation state denoted simply as ?. For the prior over $\langle o, w\rangle$ state pairs we assume a uniform distribution over the three possible observations and a uniform conditional distribution over world states given the ignorance observation state. We follow standard assumptions regarding literal compositional semantics in assigning identical unrefined literal meanings to “some” and “some or all” in the semantic lexicon. However, the more prolix “some or all” is more costly than both “some” and “all”, which are of equal cost.

Following our core assumptions laid out in Section 4.3, the set of possible lexica generated under lexical uncertainty involves all possible refinements of the meaning of each utterance: “all” cannot be further refined, but “some” and “some or all” each have three possible refinements (to $\{\forall\}$, $\{\exists\neg\forall\}$, or $\{\forall, \exists\neg\forall\}$), giving us nine lexica in total. Also following our core assumptions, each possible lexicon includes the null utterance u_{null} with maximally general meaning $[\![u_{null}]\!] = \{\exists\neg\forall, \forall\}$ and substantially higher cost than any other utterance; here we specify that cost to be $c(u_{null}) = 4$.

Figure 7 shows the results of the lexical uncertainty model under these assumptions, with greedy rationality parameter $\lambda = 4$.[7] (We chose the above parameter values to make the model's qualitative behavior easy to visualize, but the fundamental ignorance-implicature result seen here is robust across specifications of the prior probabilities, "greedy" rationality parameter, and utterance costs, so long as $c(\text{all}) = c(\text{some}) < c(\text{some or all}) < c(u_{null})$.) The key to understanding how the ignorance implicature arises lies in the S_1 matrices for lexica $\mathcal{L}_3$ and $\mathcal{L}_7$. In each of these lexica, one of *some* and *some or all* has been refined to mean only $\exists\neg\forall$, while the other remains unrefined. For a speaker whose observation state is ignorance, an utterance with a refined meaning has infinitely negative expected utility and can never be used; hence, this speaker near-categorically selects the unrefined utterance (*some* in $\mathcal{L}_3$, *some or all* in $\mathcal{L}_7$; the null utterance being ruled out due to its higher cost in both cases). But crucially, while in $\mathcal{L}_7$ the informed speaker who has observed $\exists\neg\forall$ prefers the refined utterance "some", in $\mathcal{L}_3$ that speaker prefers the *unrefined* utterance—again "some"—due to its lower cost. This asymmetry leads to an asymmetry in the marginalizing listener L_1, for whom the association with $\exists\neg\forall$ is crucially stronger for "some" than for "some or all". Further rounds of pragmatic inference strengthen the former association, which in turn drives an ignorance interpretation of "some or all" through the now-familiar mechanics that give rise to scalar implicature.

5 Compositionality

In the previous section we introduced the lexical uncertainty extension of the rational speech-act model, which surmounted a general class of challenges: explaining why two utterances with identical literal content but different form complexity receive different interpretations. In each case, lexical uncertainty led to an alignment between utterances' formal complexity and some kind of markedness of the interpretations they receive. These analyses hinged on introducing a set of refined lexica, Λ, and allowing the pragmatic reasoner to infer which lexicon from this set the speaker was using. We described how Λ could be canonically derived from a base semantic lexicon $\mathcal{L}_S$ as the set of all refined sentence meanings suitably restricted and augmented to make the model well-defined. However, there was a choice implicitly in this setup: should refinements be considered at the level of sentences, after composition has constructed meanings from lexical entries, or should refinements be considered at the level of single lexical entries, and be followed by compositional construction of sentence meaning? Our previous process, enrichment of whole sentences, operated after composition; in this section we consider an alternative, lexical enrichment, which operates before composition. In the examples we have considered so far, sentence meanings were simple enough that this choice would have little effect; as we will show below the two approaches can diverge in interesting ways for more complex sentences.

In order to generalize the previous approach to enrichment from full sentences to lexical entries of more complex types we need an extended notion of refinement. While it is beyond the scope of this paper, one could adopt the generalized notion of entailment from natural logics and then define a refinement of a lexical entry as another term of the same type that entails the original entry. The set of lexica Λ could then be derived, as before, as the set of all lexicons that can be derived from

[7] Note that interpretations in listener functions L_i are given as observation states, not pairs of observation and world states. This is a presentational shorthand; the full listener functions $L_0(o,w|u,\mathcal{L})$ and $L_i(o,w|u)$ can always be recovered by multiplying the posterior distribution on observations by the conditional distribution $P(w|o)$.

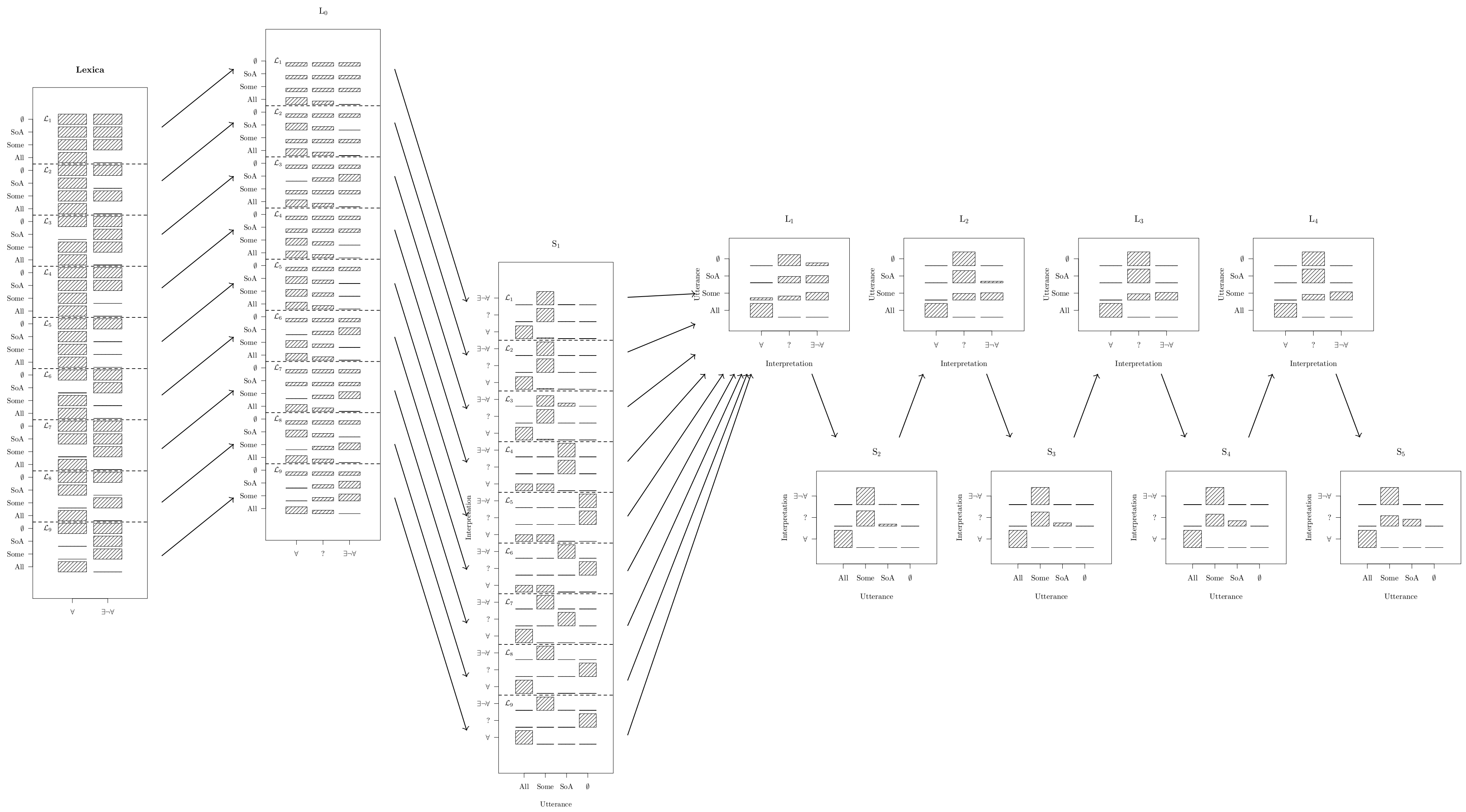


Figure 7: Markedness

$\mathcal{L}_S$ by refinement. Sentence meanings would then be derived from (refined) lexical meanings by ordinary compositional mechanisms. In this paper, we will only consider refinements of Boolean-typed lexical items. As before, we must impose certain restrictions on these refinements to ensure that the model will be well defined. The necessary restrictions are the same as in Section 4.3. Our previous solution for restriction 2 carries over: we may extend each lexicon with a trivial u_{null}. Restriction 1 is more subtle than before. We must still guarantee that the literal listener can interpret any utterance. Simply restricting that the lexical entries be assigned non-contradictory refinements in not enough, as composition can arrive an contradictions (e.g. "A and not A"). There are various options available to solve this problem[8]; below we will simply restrict our attention to composition by disjunction, where it is sufficient to require that individual lexical items are non-contradictory.

We first motivate the need to consider composition of enriched lexical entries by describing a class of implicatures that pose trouble for our approach so far. We then describe the lexical enrichment procedure for the case of Boolean composition and show that it can explain these (and other) cases of pragmatic enrichment.

5.1 Implicatures from non-convex disjunctive expressions

We have thus far explored two subtle cases of implicatures that break the symmetry between semantically equivalent utterances. The first example was that of M-implicatures such as the difference in interpretation between *Sue smiled* and *The corners of Sue's lips turned slightly upwards* (Levinson, 2000), where the relevant notion of markedness is the prior probability of the meaning: ordinary smiles are more common than smirks and grimaces. The second example was that of ignorance implicatures for disjunctions such as *some or all*, in which the relevant notion of markedness is the degree of speaker ignorance about the world state: the more complex utterance is interpreted as indicating a greater degree of speaker ignorance. However, there are even more challenging cases than these: cases in which non-atomic utterances with identical literal content *and* identical formal complexity receive systematically different interpretations. A general class of these cases can be constructed from entailment scales containing more than two items, by creating a disjunction out of two non-adjacent terms on the scale:

(i) Context: A and B are visiting a resort but are frustrated with the temperature of the springs at the resort they want to bathe in.
A: The springs in this resort are always warm or scalding. [Understood meaning: *but never hot.*]

(ii) Context: A is discussing with B the performance of her son, who is extremely smart but blows off some classes, depending on how he likes the teacher.
A: My son's performance in next semester's math class will be adequate or stellar. [Understood meaning: *but not good.*]

(iii) Context: there are four people in a dance class, and at the beginning of each class, the

[8] For instance, we could add a world state w_{err} which has non-zero weight if and only if all other states have zero weight. Since $P(w_{err}|o)=0$ for any observation o, the speaker will never choose an utterance which leads to the w_{err} interpretation. This mechanism is generally useful for filtering out un-interprettable compositions Goodman & Lassiter (2014).

students are paired up with a dance partner for the remainder of the class. A, who is not in the class, learns that one of the students in the class did not have a dance partner at a particular session, and encounters B.
B: Any idea how many of the students attended the class?
A: One or three of the students showed up to the class. [Understood meaning: *it wasn't the case that either exactly two students or exactly four students showed up.*]

These disjunctive expressions—*warm or scalding*, *decent or stellar*, *one or three*—pose two serious challenges for neo-Gricean theories. First, in each case there are alternatives disjunctive expressions with identical formal complexity (in the sense of having the same syntactic structure and number of words) and literal meaning under standard assumptions that the literal meanings of such expressions are lower bounds in the semantic space of the scale, but different understood meaning: *warm or hot*, *decent or good*, *one or two*.[9] It is not at all clear on a standard neo-Gricean account how these pairs of alternatives come to have different pragmatic interpretations. Second, these expressions have the property that their understood meanings are NON-CONVEX within the semantic space of the scale. This property poses a serious challenge for standard neo-Gricean accounts: since all the alternatives whose negation could be inferred through pragmatic reasoning have literal meanings that are upper bounds in the semantic space, it is unclear how the resulting pragmatically strengthened meaning of the utterance could ever be non-convex.

The basic lexical uncertainty framework developed in Section 4 does not provide an explanation for these cases, which we will call NON-CONVEX DISJUNCTIVE EXPRESSIONS. That framework can only derive differences in pragmatic interpretation on the basis of differences in literal meaning or complexity; in the current cases, the utterance pairs receive distinct interpretations despite sharing the same literal meaning and complexity. It turns out, however, that these cases can be elegantly handled by compositional lexical uncertainty. Before introducing the compositional lexical uncertainty framework, it is worth noting that alternative game-theoretic frameworks do not derive the appropriate interpretations of non-convex disjunctive expressions. While the IBR model is able to derive the distinction between *some* and *some or all*, it cannot derive the distinction between *one or two* and *one or three*.[10] The IBR model only derives different pragmatic interpretations based on differences in semantic content or cost; the version of the IBR model which derives the ignorance implicature for *some or all* relies on the difference in cost between *some* and *some or all* in its derivation. Because the utterances *one or two* and *one or three* have identical semantic content and complexity, the IBR model will assign these utterances identical interpretations.

[9] Explaining the difference in meaning between *one or three* and *one or two* is only a challenge for pragmatic theories if numerals have a lower-bound semantics; if numerals have an exact semantics, then these disjunctive utterances will receive different literal interpretations. However, this objection does not hold for non-numeric scales such as <*warm, hot, scalding*>, in which each lexical item has an uncontroversial lower-bound semantics. We will be using the numerical examples for illustrative purposes, but our claims will be equally applicable to the non-numeric examples.

[10] The IQR model does not provide an account of the difference in interpretation between "some" and "some or all." It is strictly more difficult to derive the appropriate implicatures in the current example — because there are strictly fewer asymmetries for the model to exploit — and therefore the IQR model will also not derive these implicatures.

5.2 Compositional lexical uncertainty

In this section we further specify compositional lexical uncertainty, as sketched out above, for the case of boolean atomic utterances composed by disjunction. This requires only a small change to the original lexical-uncertainty model introduced in Section 4: the standard assumption that the literal listener interprets non-atomic utterances by composition.

Assume that the base semantic lexicon $\mathcal{L}_S$ maps a set $\mathcal{U}_A$ of atomic utterances to Boolean-valued truth-functions (and maps "or" to the disjunction $\vee$, though we will suppress this in the notation below). The set of lexica Λ is derived by enrichment as before as all possible combinations of valid refinements of the utterance meanings in $\mathcal{L}_S$, each augmented with the always-true utterance u_{null}. From this we define denotations of (potentially non-atomic) utterances inductively. First, for an atomic utterance u, we define its denotation $[\![u]\!]_{\mathcal{L}}$ relative to lexicon $\mathcal{L}$ by:

$$[\![u]\!]_{\mathcal{L}}(w) = \mathcal{L}(u,w) \tag{40}$$

That is, the denotation of an atomic utterance relative to a lexicon is identical to its entry in the lexicon. The denotations of complex utterances are defined in the obvious inductive manner. For the disjunction "u_1 or u_2":

$$[\![u_1 \text{or } u_2]\!]_{\mathcal{L}}(w) = \begin{cases} 1 & \text{if } [\![u_1]\!]_{\mathcal{L}}(w) = 1 \text{ or } [\![u_2]\!]_{\mathcal{L}}(w) = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{41}$$

We could define the denotation of utterances built up from conjunctions and other Boolean connectives similarly (though with the caveat indicated above pertaining to contradictions), but won't need these for the below examples.

The literal listener now interprets utterances according to their denotations:

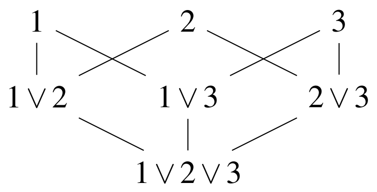
$$L_0(w,o|u,\mathcal{L}) \propto [\![u]\!]_{\mathcal{L}}(w)P(w,o) \tag{42}$$

In other words, the literal listener filters out worlds that are inconsistent with the denotation of the utterance. The definitions of the higher-order speakers and listeners are unchanged from the previous versions of the model.

5.3 Derivation of non-convex disjunctive expressions

We demonstrate the account of non-convex implicatures afforded by compositional lexical uncertainty using the running example of *one or three*, though the same account would hold for non-convex disjunctions on other scales such as *warm or scalding* and *decent or stellar*. For discursive simplicity we limit the range of the space to the integers $\{1,2,3\}$, though the account generalizes to arbitrary convex subsets of the integers. The set of ATOMIC UTTERANCES U_A and possible observation states O are, respectively:

$U_A = \{one, two, three\}$ $O =$



where the join-semilattice relationship among the seven members of O is depicted for expository convenience. The set of world states W contains what we will call only BASIC world states—in this case, 1, 2, and 3—and the mapping between world states and speaker observation states is not one-to-one. Under these circumstances, an observation state is compatible with all basic world states above it on the lattice, and observation states thus vary in the degree of speaker ignorance.

Since utterance meanings are defined as sets of world states, the literal meaning of each atomic utterance can easily be picked out as the set of world states that lie above a particular node on the join semilattice. In our running example, these nodes are $1 \vee 2 \vee 3$ for *one,* $2 \vee 3$ for *two*, and 3 for *three*. Hence we have

$$\mathcal{L}_S = \left\{ \begin{array}{ll} [\![one]\!] & = \{1,2,3\} \\ [\![two]\!] & = \{2,3\} \\ [\![three]\!] & = \{3\} \end{array} \right\}$$

for the simple indicative case.

The set of possible lexica consists of all logically possible combinations of valid refinements (i.e., non-empty subsets) of each atomic utterance's meaning. In the simple indicative case, *one* has seven possible refinements, *two* has three possible refinements, and *three* has one, hence there are twenty-one logically possible lexica, a few of which are shown below (together with denotations of complex utterances, for illustration, though they are not strictly part of the lexica):

$$\left\{ \begin{array}{ll} [\![one]\!] & = \{1,2,3\} \\ [\![two]\!] & = \{3\} \\ [\![three]\!] & = \{3\} \\ [\![one\ or\ two]\!] & = \{1,2,3\} \\ [\![two\ or\ three]\!] & = \{3\} \\ [\![one\ or\ three]\!] & = \{1,2,3\} \\ [\![one\ or\ two\ or\ three]\!] & = \{1,2,3\} \end{array} \right\} \left\{ \begin{array}{ll} [\![one]\!] & = \{3\} \\ [\![two]\!] & = \{2,3\} \\ [\![three]\!] & = \{3\} \\ [\![one\ or\ two]\!] & = \{2,3\} \\ [\![two\ or\ three]\!] & = \{2,3\} \\ [\![one\ or\ three]\!] & = \{3\} \\ [\![one\ or\ two\ or\ three]\!] & = \{2,3\} \end{array} \right\} \left\{ \begin{array}{ll} [\![one]\!] & = \{1\} \\ [\![two]\!] & = \{2\} \\ [\![three]\!] & = \{3\} \\ [\![one\ or\ two]\!] & = \{1,2\} \\ [\![two\ or\ three]\!] & = \{2,3\} \\ [\![one\ or\ three]\!] & = \{1,3\} \\ [\![one\ or\ two\ or\ three]\!] & = \{1,2,3\} \end{array} \right\}$$

To show how this account correctly derives understood meanings for non-convex disjunctive utterances, we need to complete the model specification by choosing utterance costs and prior probabilities. Similar to the approach taken in Section 4.6.2, we make the minimally stipulative assumptions of (i) a uniform distribution over possible observations, (ii) a uniform conditional distribution for each observation over all worlds compatible with that observation; and (iii) a constant, additive increase in utterance cost for each disjunct added to the utterance. We set the cost per disjunct arbitrarily at 0.05 and set λ to 5, though our qualitative results are robust to precise choices of (i–iii) and of λ.

Here we examine in some detail how the model correctly accounts for interpretations of non-convex disjunctive expressions in the simple indicative case. Even in this case there are 21 lexica, which makes complete visual depiction unwieldy; for simplicity, we focus on the twelve lexica in which the denotation of *one* has not been refined to exclude 1, because it is in this subset of lexica in which *one* has already been distinguished from *two* and we can thus focus on the inferential dynamics leading to different interpretations for *one or two* versus *one or three*. Figure 8 shows the behavior of this pragmatic reasoning system. The three leftmost panels show the twelve lexica and the resulting literal-listener L_0 and first-level speaker S_1 distributions respectively; the three rightmost panels show the marginalizing listener L_1 and the subsequent speaker and listener S_2 and L_2 respectively; by the L_2 level, pragmatic inference has led both atomic and disjunctive utterances to be near-categorically associated with interpretations such that each atomic term in an utterance

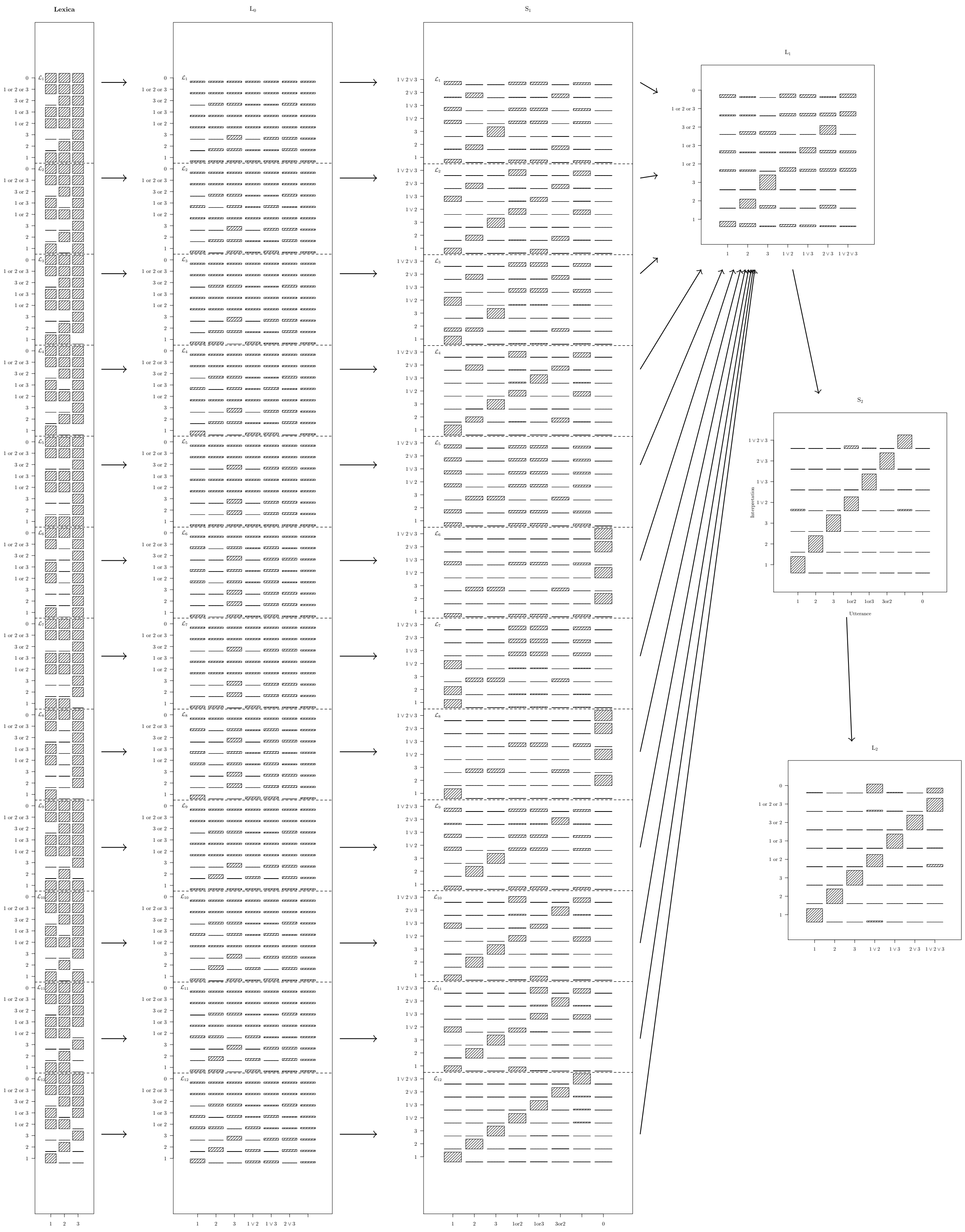


Figure 8: Non-convex disjunction, for uniform marginal distribution $P(O)$, uniform conditional distributions $P(W|O)$, cost per disjunct of 0.05, and $\lambda = 5$. Only lexica (and L_0 and S_1 distributions) in which the refined meaning of *one* contains the world state 1 are shown.

has an exact meaning at the lower bound of the term's unrefined meaning (and such that disjunctive utterances are thus disjunctions of exact meanings). The key to understanding why this set of interpretations is obtained can be found in the asymmetries among possible refinements of atomic terms in the lexica. Observe that under lexical uncertainty both *two* and *three* can have refined meanings of $\{3\}$; but whereas *three* MUST have this meaning, *two* has other possible meanings as well ($\{2\}$ and $\{2,3\}$). Consequently, the set of lexica in which *one or two* has $\{1 \vee 3\}$ as its meaning ($\mathcal{L}_6$ and $\mathcal{L}_8$) is a strict subset of the set of lexica in which *one or three* has that meaning (which also includes $\mathcal{L}_2$, $\mathcal{L}_4$, $\mathcal{L}_{10}$, and $\mathcal{L}_{12}$). Pragmatic inference leads to a strong preference at the S_1 level in the latter four lexica for expressing observation state $\{1 \vee 3\}$ with *one or three*, even in $\mathcal{L}_4$ and $\mathcal{L}_{10}$ where that observation state is compatible with the utterance *one or two*. Furthermore, there are no lexica in which the reverse preference for expressing $\{1 \vee 3\}$ with *one or two* is present at the S_1 level. This asymmetry leads to a weak association between *one or three* and $\{1 \vee 3\}$ for the marginalizing L_1 listener, an association which is strengthened through further pragmatic inference.

5.4 *Some or all* ignorance implicatures with compositional lexical uncertainty

For completeness, we briefly revisit the ignorance implicatures of *some or all* originally covered in Section 4.6, now within the framework of compositional lexical uncertainty. In short, compositional lexical uncertainty derives ignorance implicature for *some or all* for similar reasons that it derives interpretations for the more difficult cases of non-convex disjunctive expressions: there are lexica in which *some* is refined to mean $\{\exists\neg\forall\}$, but no lexica in which *some or all* can be refined to have this meaning. This asymmetry leads to a weak association for the marginalizing L_1 listener between *some* and $\exists\neg\forall$ and between *some or all* and the ? ignorant-speaker observation state. Further pragmatic inference strengthens this association (S_2 and L_2).[11]

6 Discussion

We have discussed a sequence of increasingly complex pragmatic phenomena, and described a corresponding sequence of probabilistic models to account for these phenomena. The first, and simplest, phenomena discussed were specificity implicatures, a generalization of scalar implicatures: the inference that less (contextually) specific utterances imply the negation of more specific utterances. These implicatures can be derived by the Rational Speech Acts model (Goodman & Stuhlmüller, 2013), a model of recursive social reasoning. This model, which is closely related to previous game-theoretic models of pragmatics, represents the participants in a conversation as rational agents who share the goal of communicating information with each other; the model's assumptions closely track those of traditional Gricean accounts of pragmatic reasoning. In addition

[11] It is worth remarking that this asymmetry resulting from the constraints across denotations of utterances imposed by compositional lexical uncertainty is strong enough to derive the empirically observed interpretations and associated ignorance implicatures of disjunctive expressions even without any differences in utterance costs. Thus compositional lexical uncertainty can be viewed as a fully-fledged alternative to the "ignorance as a marked state" view of the basic ignorance implicatures of Section 4.6.

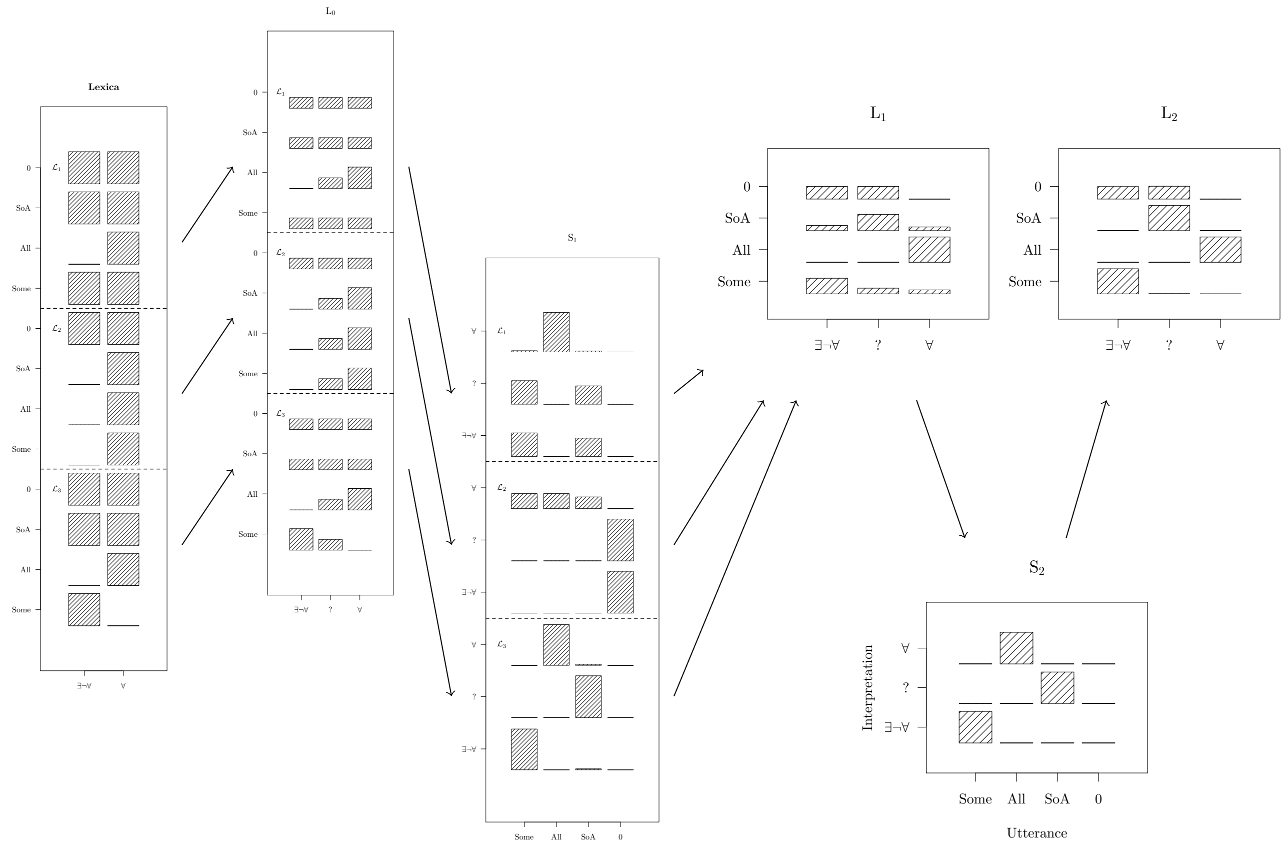


Figure 9: *Some or all* ignorance implicature under compositional lexical uncertainty.

to using this model to derive specificity implicatures, we showed that it can be used to provide a solution to the symmetry problem for scalar implicatures.

We next turned to M-implicatures, in which complex utterances are assigned low probability interpretations, while simpler but semantically equivalent utterances are assigned higher probability interpretations. We showed that the rational speech acts model does not derive these implicatures. The reasons for this failure are related to the multiple equilibrium problem for signaling games, a general barrier to deriving M-implicatures in game-theoretic models. In order to account for these implicatures, we introduced lexical uncertainty, according to which the participants in a conversation have uncertainty about the semantic content of their utterances. We showed that, with this technique, the participants in a conversation derive M-implicatures by using pragmatic inference to resolve the semantic content of potential utterances.

Both specificity implicatures and M-implicatures can be derived given the assumption that the speaker is fully knowledgeable about the true world state (at the relevant degree of granularity). Following our derivations of these inferences, we examined several classes of inferences which require this knowledgeability assumption to be relaxed. The first of these was the ignorance implicature associated with the expression *some or all*. The rational speech acts model fails to derive this implicature for reasons which are nearly identical to its failure to derive M-implicatures. Surprisingly, we showed that the lexical uncertainty model does derive this implicature: according to this model, the ignorance implicature arises because of the greater complexity of *some or all* relative to its alternative *some*. This suggests that the lexical uncertainty model captures a generalized notion of markedness, according to which complex utterances received marked interpretations, and where markedness may indicate low probability, ignorance, and possibly other features.

We finally examined a general class of Hurford-violating embedded implicatures, in which equally complex — and semantically equivalent — utterances such as *one or two* and *one or three* are assigned distinct interpretations. Because the basic lexical uncertainty model can only derive distinct pragmatic interpretations for a pair of utterances by leveraging either differences in semantic content or complexity, it is unable to derive this class of implicature. We therefore considered extending the framework to compositional lexical uncertainty, which respects the compositional structure of utterances. By performing inference on the semantic content of sub-sentential expressions, this model derives the class of embedded implicatures we considered, and gives a richer role to compositional structure.

In the remainder, we will discuss several conceptual questions about the modeling framework proposed in this paper, and will note some further applications of these ideas.

6.1 Semantic underspecification as uncertainty

The core conceptual and technical idea involved in lexical uncertainty is to model the speaker and listener as uncertain about the semantic content of their language's utterances. In this paper we have assumed that the "base" lexicon fully specifies the semantic content of each lexical item; semantic uncertainty, comes from uncertainty about the specific refinement of this base lexicon that is being used. There is an alternative way of formalizing semantic uncertainty, which preserves all of the modeling results presented in this paper but which pushes the semantic uncertainty into the lexical representations themselves.

This alternative formalization uses *semantic free variables* as a locus of uncertainty [12]. These semantic free variables are used to assign underspecified semantic content to utterances. This is done in the usual manner (Lewis, 1970; Montague, 1973): certain variables in a lexical entry are left un-bound; the semantic content of a lexical item is fully specified once all of the free variables in its lexical entry have been assigned values based on context. Semantic uncertainty can be represented as uncertainty about the values of the relevant variables. For example, we might assign the adjective "tall" the lexical entry $\lambda x \lambda y.\text{height}(y) > x$, where the value of the threshold variable x is left unspecified in the lexicon (Lassiter & Goodman, 2013). The value of this variable is inferred during pragmatic inference, jointly with the world state, in a manner identical to the lexical inference procedure described in this paper.

Though lexical uncertainty and the semantic free variable technique use distinct representations, it is fairly straightforward to show that they are equivalent up to parameterization. We informally sketch this equivalence: Given a set of underspecified lexical items, we can fully fix the semantic interpretation of these items by assigning values to all of their semantic free variables. Thus, a complete assignment of values to the lexical items' semantic free variables can be represented by a lexicon which assigns each lexical item a fully specified semantic content. The joint distribution over the values of semantic free variables therefore can be represented by a distribution

[12] Lexical uncertainty, as first proposed in Bergen et al. (2012), assumes that the lexicon fully specifies the semantic content of each lexical item. The reinterpretation in terms of semantic free variables was proposed in Lassiter & Goodman (2013).

over lexicons, and it follows almost immediately that lexical uncertainty can simulate semantic free variables. (Though note that this distribution over lexica is more general than the all-refinements prior we have favored in the examples worked here.)

It is similarly straightforward to use semantic free variables to simulate lexical uncertainty. For instance, take a lexical item with entry $\lambda y.C(y)$, where C is a fully specified predicate. We can create an abstracted version of this entry by conjoining a free predicate variable to it: $\lambda F \lambda y.C(y) \wedge F(y)$. This F variable will represent the enriched semantic content that this lexical item will convey, in addition to the invariant semantic content conveyed by the predicate C. We can represent a distribution over lexical entries assigned to this lexical item by a suitable distribution over values to assign the free variable of the abstracted lexical entry. Therefore we can represent a distribution over lexicons by abstracting all of the lexical entries in this manner, and defining suitable distribution over values to assign to the variables in these lexical entries.

Hence there is no substantive commitment involved in using lexical uncertainty rather than semantic free variables, or vice-versa. Nonetheless, the choice of representation will have an effect on how simple it is to describe certain models, and what prior distributions appear natural. For example, uncertainty about the meaning of “tall” is naturally represented as uncertainty about the height threshold that an object is required to meet. It is less natural to represent this using refinements of a fully specified base meaning; one can define the required distribution over refinements of a fully permissive meaning but lexical uncertainty does not itself explain why this distribution should be preferred over the many other possible distributions over enrichments of “tall.” In contrast, lexical uncertainty provides a natural representation for deriving M-implicatures, such as the difference in interpretation between “John can finish the homework” and “John has the ability to finish the homework.”

An important implication of the free-variable interpretation of lexical uncertainty is that any case previously identified as containing semantic underspecification potentially supports the kinds of complex pragmatic interactions described here. That is, our formalization of pragmatic inference formalizes the pragmatic resolution of contextual variables that have been used since the dawn of compositional semantics (Lewis, 1970; Montague, 1973). Further research will be needed to determine if this is the right theory of inference for all free variables.

6.2 Utterance cost and complexity

The notion of utterance cost plays an important role in the explanations of a number of phenomena discussed in this paper. The proposed solution to the symmetry problem relies on assigning non-salient alternatives a higher cost than salient alternatives; the derivation of M-implicatures requires a cost asymmetry between the utterance that will be assigned a high-probability meaning and the one that will be assigned a low-probability meaning; and the more general treatment of markedness requires that utterances receiving marked interpretations be more costly.

One interpretation of the cost parameter in our models is that it represents how much *effort* is required for the speaker to convey an utterance. This effort may reflect the length of the utterance (in, e.g., syllables); the difficulty of correctly pronouncing it; the amount of energy required to produce the sounds required for the utterance; the effort to recall appropriate words from memory; or still other possible factors. An interpretation of the cost parameter in this manner constitutes a theory of how the speaker chooses utterances, as well as a theory of how *the listener* believes the speaker chooses utterances.

An additional feature of utterances that may effect utterance choice, one which is less clearly related to effort, is the utterance complexity under the speaker's theory of their language. That is, the speaker may be less likely to use a particular utterance, not necessarily because it is difficult to say, but because it is a complex utterance according to their grammar. For example, the speaker may be unlikely to use the locative-inversion construction, "Onto the table jumped the cat," even though by all appearances it is no more difficult to say than, "The cat jumped onto the table"; this is attested in the corpus frequencies for these constructions, where the locative inversion is much less common. A theory of how the speaker chooses utterances should thus be sensitive to some notion of linguistic complexity. It is possible that effort indeed tracks complexity (for instance a resource-rational analysis might predict that language is processed in such a way that more common utterances are easier to access and produce). Or it may be that this is an orthogonal aspect of speaker utility that must be encoded in the utterance cost. Fortunately it is straightforward to represent linguistic complexity in our models (e.g. by adding log of the probability of the utterance under a PCFG to the utility), and to derive exactly the same predictions starting from differences in complexity rather than differences in difficulty. Future work will be required to clarify the specific form and nature of the cost model.

6.3 Embedded implicatures

We have used lexical uncertainty to derive implicatures which arise from Hurford-violating disjunctions. We have focussed on these implicatures because they pose a particularly strong challenge for Gricean/game-theoretic models of pragmatics. In particular, it has been argued that they provide evidence that certain implicatures must be computed locally in the grammar, through the use of an exhaustivity operator (Chierchia et al., 2012). The arguments for this position are closely related to the previously discussed challenges in deriving these implicatures using game-theoretic models: A Hurford-violating disjunction is semantically equivalent to one of its disjuncts. As a result, pragmatic theories which posit only global pragmatic computations will not be able to straightforwardly derive the implicatures associated with these disjunctions, because these theories typically rely on differences in semantic content between whole utterances to derive pragmatic inferences. These embedded implicatures differ in a crucial way from many others discussed in the literature: in these other cases, the implicature-generating utterance is semantically distinct from its relevant alternatives (Chierchia, 2006; Fox, 2007). For example, the sentence *Kai had broccoli or some of the peas last night* has a distinct semantic interpretation from its nearby alternatives, and in particular, from any alternative which has a distinct set of implicatures. The argument that global approaches to pragmatic reasoning cannot derive these implicatures is therefore much less straightforward for these utterances; the most one can typically show is that a specific model of pragmatic reasoning does not derive the implicatures in question. Indeed, it has been argued that many of these implicatures can be derived by global pragmatic reasoning (Russell, 2006; Sauerland, 2004). The lexical uncertainty approach also predicts many of these weaker, but more discussed, embedded implicatures, though we will not give details of these derivations here. The success of lexical uncertainty in deriving the Hurford-violating embedded implicatures, which pose the greatest challenge, provides an encouraging piece of evidence that the general class of probabilistic, social-reasoning-based models can explain the empirical phenomena of embedded implicatures.

7 Conclusion

In this paper we have explored a series of probabilistic models of pragmatic inference. The initial Rational Speech Acts model (Goodman & Stuhlmüller, 2013) straightforwardly captures the Gricean imperatives that the speaker be informative but brief, and that the listener interpret utterances accordingly. This model predicts a variety of pragmatic enrichments, but fails to derive M-implicatures and several other implicature patterns. We have thus moved beyond the traditional Gricean framework to consider pragmatic reasoning over lexical entries—inferring the "literal meaning" itself. In this framework the impetus driving pragmatic enrichment is not only alternative utterances, but alternative semantic refinements. Thus uncertain or underspecified meanings have the opportunity to contribute directly to pragmatic inference. We showed that this *lexical uncertainty* mechanism was able to derive M-implicatures, Hurford-violating embedded implicatures, and a host of other phenomena.

References

Bergen, Leon, Noah D Goodman & Roger Levy. 2012. That's what she (could have) said: How alternative utterances affect language use. In *Proceedings of the thirty-fourth annual conference of the cognitive science society*, .

Chierchia, Gennaro. 2006. Broaden your views: Implicatures of domain widening and the "logicality" of language. *Linguistic inquiry* 37(4). 535–590.

Chierchia, Gennaro, Danny Fox & Benjamin Spector. 2012. Scalar implicature as a grammatical phenomenon. In Klaus von Heusinger, Claudia Maienborn & Paul Portner (eds.), *Semantics: An international handbook of natural language meaning*, vol. 3, chap. 87. Berlin: Mouton de Gruyter.

Cho, In-Koo & David M Kreps. 1987. Signaling games and stable equilibria. *The Quarterly Journal of Economics* 102(2). 179–221.

Clark, Herbert H. 1996. *Using language*. Cambridge University Press.

De Jaegher, Kris. 2008. The evolution of horn's rule. *Journal of Economic Methodology* 15(3). 275–284.

Degen, Judith, Michael Franke & Gerhard Jäger. 2013. Cost-based pragmatic inference about referential expressions. In Markus Knauff, Michael Pauen andNatalie Sebanz & Ipke Wachsmuth (eds.), *Proceedings of the annual meeting of the cognitive science society*, 376–381.

Fox, Danny. 2007. Free choice and the theory of scalar implicatures. In Uli Sauerland & Penka Stateva (eds.), *Presupposition and implicature in compositional semantics*, 71–120. Basingstoke: Palgrave Macmillan.

Fox, Danny. 2014. Cancelling the Maxim of Quantity: Another challenge for a Gricean theory of scalar implicatures. *Semantics and Pragmatics* 7(5). 1–20.

Fox, Danny & Roni Katzir. 2011. On the characterization of alternatives. *Natural Language Semantics* 19(1). 87–107.

Frank, Michael C & Noah D Goodman. 2012. Predicting pragmatic reasoning in language games. *Science* 336(6084). 998–998.

Franke, Michael. 2009. Signal to act: Game theory in pragmatics .

Franke, Michael & Gerhard Jäger. 2013. Pragmatic back-and-forth reasoning. *manuscript, Amsterdam & Tübingen* .

Fudenberg, Drew & Jean Tirole. 1991. Game theory. *Cambridge, MA* .

Gazdar, Gerald. 1979. *Pragmatics: Implicature, presupposition, and logical form*. Academic Press New York.

Goodman, Noah D & Daniel Lassiter. 2014. Probabilistic semantics and pragmatics: Uncertainty in language and thought. *Handbook of Contemporary Semantic Theory, Wiley-Blackwell* .

Goodman, Noah D & Andreas Stuhlmüller. 2013. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science* 5(1). 173–184.

Grice, H Paul. 1975. Logic and conversation. *1975* 41–58.

Harsanyi, John C. 1967. Games with incomplete information played by ”bayesian” players, i-iii. part i. the basic model. *Management Science* 14(3). 159–182.

Hirschberg, Julia Linn Bell. 1985. *A theory of scalar implicature*. University of Pennsylvania.

Horn, Laurence. 1984. Toward a new taxonomy for pragmatic inference: Q-based and r-based implicature. *Meaning, form, and use in context* 42.

Hurford, James R. 1974. Exclusive or inclusive disjunction. *Foundations of Language* 409–411.

Jäger, Gerhard. 2012. Game theory in semantics and pragmatics. In Claudia Maienborn, Paul Portner & Klaus von Heusinger (eds.), *Semantics: An international handbook of natural language meaning*, De Gruyter Mouton.

Lassiter, Daniel & Noah D Goodman. 2013. Context, scale structure, and statistics in the interpretation of positive-form adjectives. In *Proceedings of salt*, .

Levinson, Stephen C. 2000. *Presumptive meanings: The theory of generalized conversational implicature*. MIT Press.

Lewis, David. 1969. *Convention: A philosophical study* 80. Harvard University Press.

Lewis, David. 1970. General semantics. *Synthese* 22(1). 18–67.

Lewis, Richard L. & Shravan Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science* 29. 1–45.

Meyer, Marie-Christine. 2013. *Ignorance and grammar*: Massachusetts Institute of Technology dissertation.

Montague, Richard. 1973. The proper treatment of quantification in ordinary English. In Jaakko Hintikka, Julius Matthew Emil Moravcisk & Patrick Suppes (eds.), *Approaches to natural language*, 221–242. Dordrecht: D. Reidel.

Myerson, Roger B. 2013. *Game theory: analysis of conflict*. Harvard university press.

Nash, John F et al. 1950. Equilibrium points in n-person games. *Proceedings of the national academy of sciences* 36(1). 48–49.

Parikh, Prashant. 2000. Communication, meaning, and interpretation. *Linguistics and Philosophy* 23(2). 185–212.

Rabin, Matthew. 1990. Communication between rational agents. *Journal of Economic Theory* 51(1). 144–170.

Rothschild, Daniel. 2013. Game theory and scalar implicatures. *Philosophical Perspectives* 27(1). 438–478.

Russell, Benjamin. 2006. Against grammatical computation of scalar implicatures. *Journal of semantics* 23(4). 361–382.

Russell, Benjamin. 2012. *Probabilistic reasoning and the computation of scalar implicatures*: Brown University dissertation.

Sauerland, Uli. 2004. Scalar implicatures in complex sentences. *Linguistics and philosophy* 27(3). 367–391.

Smith, Nathaniel J. & Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition* 128(3). 302–319. http://dx.doi.org/10.1016/j.cognition.2013.02.013.

Sperber, Dan & Deidre Wilson. 1986. *Relevance: Communication and cognition*, vol. 142. Cambridge, MA: Harvard University Press.

Stalnaker, Robert. 1978. Assertion. *Syntax and Semantics (New York Academic Press)* 9. 315–332.

Sutton, Richard S & Andrew G Barto. 1998. *Reinforcement learning: An introduction*. Cambridge Univ Press.

Tenenbaum, Joshua B., Charles Kemp, Thomas L. Griffiths & Noah Goodman. 2011. How to grow a mind: statistics, structure, and abstraction. *Science* 331. 1279–1285.

Van Rooy, Robert. 2004. Signalling games select horn strategies. *Linguistics and Philosophy* 27(4). 493–527.

A Experimental validation of ignorance implicature

Here we will describe an experimental evaluation of the linguistic judgments discussed in Section 4.6. For ease of exposition, we will reproduce the examples from that section here:

(i) Some or all of the students passed the test.

(ii) Some of the students passed the test.

The experiment evaluated two claims about the interpretation of example (i). The first claim is that while example (ii) implicates that not all of the students passed the test, example (i) does not carry this implicature. The second claim is that this example carries an ignorance implicature: it implicates that the speaker does not know whether all of the students passed.

A.1 Methods

Participants Thirty participants were recruited from Amazon's Mechanical Turk, a web-based crowdsourcing platform. They were provided with a small amount of compensation for participating in the experiment.

Materials We constructed six items of the following form:

> Letters to Laura's company almost always have checks inside. Today Laura received 10 letters. She may or may not have had time to check all of the letters to see if they have checks. You call Laura and ask her how many of the letters have checks inside. She says, "{Some/Some or all} of the letters have checks inside."

The name of the speaker (e.g. "Laura") and the type of object being observed (e.g. checks inside letters) were varied between items. The speaker's utterance was varied within items, giving two conditions for each item, "Some" and "Some or all." Each participant was shown every item in a randomly assigned condition.

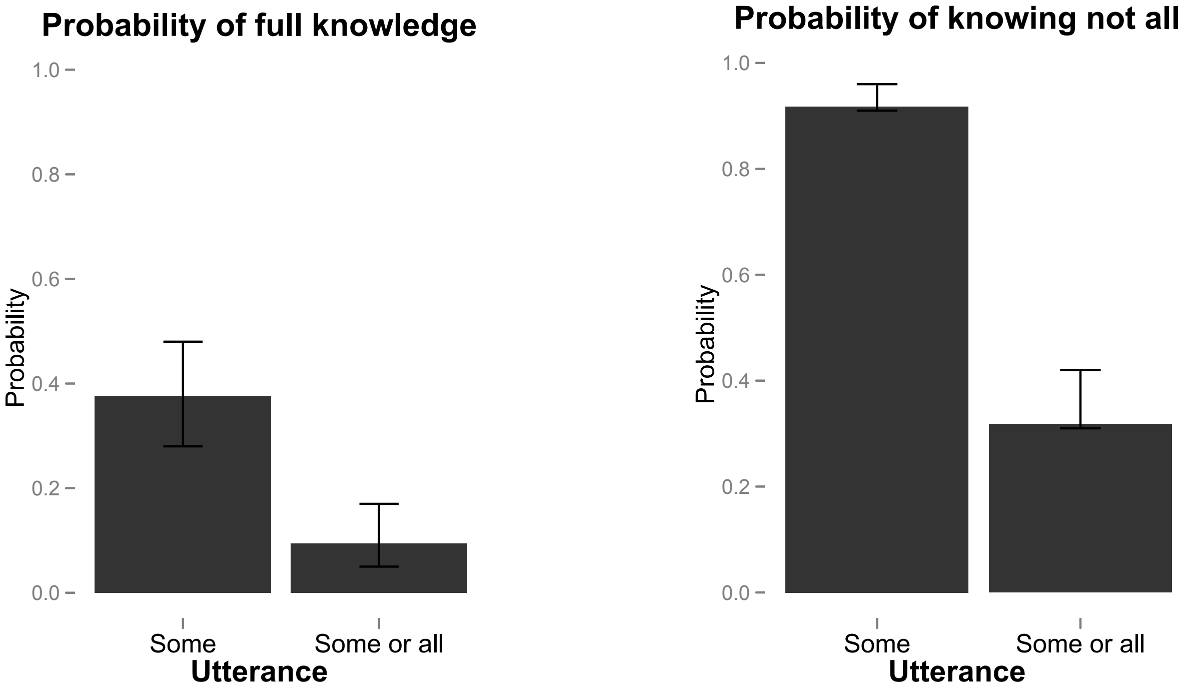
After reading an item, participants were asked two questions:

A: *How many letters did Laura look inside?*

B: *Of the letters that Laura looked inside, how many had checks in them?*

Question **A** was used to assess whether the speaker knows *all*, which in this example would mean that Laura knows that all of the letters have checks inside of them. This question assesses whether the speaker meets a necessary condition on knowing *all*. If, for example, Laura has not looked inside each letter, then she cannot know that all of the letters have checks inside. Question **B** was used to assess whether the speaker knows *not all*, which in this example would mean that Laura looked inside letters which did not have checks in them. If the numerical response to the first question exceeds the response to the second question, then Laura knows that not all of the letters have checks in them.

A.2 Results



(a) $P(\mathbf{A} = 10)$ as a function of the speaker's utterance. Error bars are 95% confidence intervals.

(b) $P(\mathbf{A} > \mathbf{B})$ as a function of the speaker's utterance.

Figure 10: Interpretation of the two speaker utterances.

We first analyzed the effect of the speaker's utterance on judgments of whether the speaker observed the full world state, as measured by responses to Question **A**. In particular, we analyzed the effect on the probability that the speaker examined all 10 objects, which we denote by $P(\mathbf{A} = 10)$. This analysis was performed using a logistic mixed-effects model, with random intercepts and slopes for items and participants. Responses in the "Some or all" condition were significantly less likely to indicate that the speaker examined all 10 objects than in the “Some” condition ($\beta = -5.81$; $t = -2.61$; $p < 0.01$). This result is shown in Figure 10a.

We next analyzed the effect of the speaker’s utterance on judgments of whether the speaker knows *not all*. This was measured using the probability that the number of total observations (as measured by the response to Question **A**) was greater than the number of positive observations (as measured by Question **B**). This probability is denoted by $P(\mathbf{A} > \mathbf{B})$. The analysis was performed using a logistic mixed-effects model, with random intercepts for participants, and random intercepts and slopes for items.[13] Responses in the “Some or all” condition were significantly less likely to indicate that $\mathbf{A} > \mathbf{B}$ than those in the “Some” condition ($\beta = -4.73$; $t = -7.22$; $p < 0.001$). This result is shown in Figure 10b.

These results provide evidence for the two claims about the interpretation of “Some or all.” First, while “Some” carries a specificity implicature, and indicates that the speaker knows *not all*, “Some or all” does not carry this implicature, and instead indicates that the speaker does not know *not all*. Second, “Some or all” indicates that the speaker also does not know *all*. Together, this provides evidence that “Some or all” carries an ignorance implicature, providing information that the speaker does not know the full state of the world.

B Two incorrect definitions of lexical uncertainty

In Section 4.3, we defined lexical uncertainty, and in Section 4.5, we used this technique to derive M-implicatures. The definition of lexical uncertainty contains several subtle assumptions about the speaker’s and listener’s knowledge of the lexicon. In this section, we will examine these assumptions in more detail, and will demonstrate that two alternative definitions which violate these assumptions fail to derive M-implicatures.

Consider the definition of lexical uncertainty in Equations 26, 28, and 29. These equations can be taken to represent the following set of claims about the speaker’s and listener’s beliefs: a) the listener L_1 believes that the speaker S_1 believes that the listener L_0 is using a particular lexicon; b) the listener L_1 believes that the speaker S_1 is certain about which lexicon the listener L_0 is using; and c) the listener L_1 is uncertain about which lexicon is being used by S_1 and L_0.

This description of the model highlights one of its non-intuitive features: the listener L_1 is uncertain about the lexicon, but believes that the less sophisticated agents S_1 and L_0 are certain about it. The description also suggests two natural alternatives to this model which one might consider. Both of these alternatives involve removing the lexical uncertainty from listener L_1 and placing it elsewhere. Under *the L_0-uncertainty model*, the literal listener L_0 is defined as being uncertain about the lexicon. Under *the S_1-uncertainty model*, the speaker S_1 is defined as being uncertain about the lexicon.

We will first show that the L_0-uncertainty model does not derive M-implicatures. The definition of this alternative model requires a single modification to the rational speech acts model from Section 2. The literal listener is now defined as being uncertain about which lexicon to use for interpreting utterances:

$$L_0^{unc}(o,w|u) \propto \sum_{\mathcal{L}'} P(\mathcal{L}')P(o,w)\mathcal{L}'(u,w) \tag{43}$$

[13] The model which included random slopes for participants did not converge.

Whereas the rational speech acts model uses a fixed lexicon $\mathcal{L}$ for interpretation, the literal listener in this model interprets utterances by averaging over lexica. The distribution $P(\mathcal{L})$ over lexica is defined to be the same as in Section 4.3.

Lemma 3. *For every distribution $P(\mathcal{L})$ over lexica, there exists a lexicon $\mathcal{L}_P$ such that $L_0^{unc}(\cdot|u) = L_0(\cdot|u,\mathcal{L}_P)$.*

Proof. Let $P(\mathcal{L})$ be the distribution over lexica in equation 43. Define the lexicon $\mathcal{L}_P$ as follows:

$$\mathcal{L}_P(u,w) = \sum_{\mathcal{L}} P(\mathcal{L})\mathcal{L}(u,w) \tag{44}$$

Then it follows from equation 43 that:

$$\begin{aligned} L_0^{unc}(o,w|u) &\propto \sum_{\mathcal{L}'} P(\mathcal{L}')P(o,w)\mathcal{L}'(u,w) && (45)\\ &= P(o,w)\sum_{\mathcal{L}'} P(\mathcal{L}')\mathcal{L}'(u,w) && (46)\\ &= P(o,w)\mathcal{L}_P(u,w) && (47)\\ &\propto L_0(o,w|u,\mathcal{L}_P) && (48)\end{aligned}$$

The last line follows by noting that this is identical to the definition of the literal listener in equation 2. Because both $L_0^{unc}(\cdot|u)$ and $L_0(\cdot|u,\mathcal{L}_P)$ define distributions, it follows that $L_0^{unc}(\cdot|u) = L_0(\cdot|u,\mathcal{L}_P)$. □

This lemma shows that the literal listener in the L_0-uncertainty model can be equivalently defined as a literal listener who is certain that the lexicon is $\mathcal{L}_P$. The listener L_0 in the new model is therefore equivalent to a listener L_0 in the rational speech acts model. Because the L_0-uncertainty model is identical to the rational speech acts model for all agents other than L_0, it follows that the L_0-uncertainty model is an instance of the rational speech acts model.

Lemma 4. *Let lexicon $\mathcal{L}_P$ be as defined in Lemma 3. Suppose u,u' are utterances that have identical interpretations according to the semantic lexicon $\mathcal{L}_S$. Then $L_0(\cdot|u,\mathcal{L}_P) = L_0(\cdot|u',\mathcal{L}_P)$.*

Proof. Let Λ be the set of lexica as defined in Section 4.3, and let $P(\mathcal{L})$ be the distribution over lexica defined there. Let $f : \Lambda \to \Lambda$ be the bijection that results from swapping the lexical entries for u and u' in each lexicon. By the definition of f, $\mathcal{L}(u,w) = f(\mathcal{L})(u',w)$ for all lexica $\mathcal{L}$ and worlds w. Because u and u' have the same interpretations in the semantic lexicon L_S, it follows that $f(\mathcal{L})$ is an admissible lexicon iff $\mathcal{L}$ is admissible. Furthermore, because $P(\mathcal{L})$ is the maximum entropy distribution over admissible lexica, $P(\mathcal{L}) = P(f(\mathcal{L}))$.

Given this bijection f,

$$\begin{aligned} L_0(o,w|u,\mathcal{L}_P) &\propto P(o,w)\mathcal{L}_P(u,w) && (49)\\ &= P(o,w)\sum_{\mathcal{L}'} P(\mathcal{L}')\mathcal{L}'(u,w) && (50)\\ &= P(o,w)\sum_{\mathcal{L}'} P(f(\mathcal{L}'))f(\mathcal{L}')(u',w) && (51)\\ &= P(o,w)\mathcal{L}_P(u',w) && (52)\\ &\propto L_0(o,w|u',\mathcal{L}_P) && (53)\end{aligned}$$

Equality between $L_0(\cdot|u,\mathcal{L}_P)$ and $L_0(\cdot|u',\mathcal{L}_P)$ follows from the fact that both define probability distributions. □

These two lemmas have established that the L_0-uncertainty model is an instance of the rational speech acts model, and that the listener L_0 interprets utterances u,u' identically if they are assigned identical semantic interpretations. Combining these results with Lemma 2, it follows that the L_0-uncertainty model does not derive M-implicatures.

We will now show that the S_1-uncertainty model does not derive M-implicatures. The definition of this model also requires a single modification to the rational speech acts model. The change comes in the definition of the utility for speaker S_1:

$$U_1(u|o) = -D_{KL}(P_o||L_a(\cdot|u)) - c(u) \tag{54}$$

where L_a is defined by:

$$L_a(\cdot|u) = \sum_{\mathcal{L}} P(\mathcal{L}) L_0(\cdot|u,\mathcal{L}) \tag{55}$$

This model represents the speaker S_1 as having uncertainty about the lexicon, and as trying to minimize the distance between their beliefs and the expected beliefs of the listener L_0. As the definition suggests, the expectation over the listener's beliefs can be represented by an average listener L_a. The distribution $P(\mathcal{L})$ over lexica is again defined to be the same as in Section 4.3.

Lemma 5. *Let utterances u,u' be assigned identical interpretations by the semantic lexicon $\mathcal{L}_S$. Then, as defined by equation 55, $L_a(\cdot|u) = L_a(\cdot|u')$.*

Proof. Let $f : \Lambda \to \Lambda$ be a bijection on the set of lexica as defined in Lemma 4. By expanding the definition of L_a, we see that:

$$\begin{aligned}
L_a(o,w|u) &= \sum_{\mathcal{L}} P(\mathcal{L}) L_0(o,w|u,\mathcal{L}) && (56)\\
&= \sum_{\mathcal{L}} P(\mathcal{L}) \frac{P(o,w)\mathcal{L}(u,w)}{Z_{u,\mathcal{L}}} && (57)\\
&= \sum_{\mathcal{L}} P(f(\mathcal{L})) \frac{P(o,w) f(\mathcal{L})(u',w)}{Z_{u',f(\mathcal{L})}} && (58)\\
&= \sum_{\mathcal{L}} P(f(\mathcal{L})) L_0(o,w|u',f(\mathcal{L})) && (59)\\
&= L_a(o,w|u') && (60)
\end{aligned}$$

The term $Z_{u,\mathcal{L}}$ is the normalizing constant for the distribution $L_0(\cdot|u,\mathcal{L})$, and the equality $Z_{u,\mathcal{L}} = Z_{u',f(\mathcal{L})}$ follows from the fact that $\mathcal{L}(u,w) = f(\mathcal{L})(u',w)$ for all lexica $\mathcal{L}$. □

This lemma establishes that if two utterances are equivalent under the semantic lexicon, then the average listener L_a will interpret them identically. For all agents more sophisticated than the average listener L_a, the S_1-uncertainty model coincides with the rational speech acts model. By Lemma 2, this is sufficient to show that the S_1-uncertainty model does not derive M-implicatures.