Semantics & Pragmatics Volume 9, Article 20: 1–91, 2016
https://doi.org/10.3765/sp.9.20

Pragmatic reasoning through semantic inference*

Leon Bergen
Massachusetts Institute of Technology

Roger Levy
University of California, San Diego

Noah Goodman
Stanford University

Submitted 2015-10-21 / Decision 2015-12-20 / Revision received 2016-02-05 / Accepted 2016-04-27 / Published 2016-04-28 / Final typesetting 2024-11-04

Abstract A number of recent proposals have used techniques from game theory and Bayesian cognitive science to formalize Gricean pragmatic reasoning (Franke 2009, Frank & Goodman 2012, 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 (Hurford 1974, Chierchia, Fox & Spector 2012). While techniques have been developed for deriving M-implicatures, Hurford-violating embedded implicatures pose a more fundamental challenge, because of basic limitations in the models' 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. We use this principle to explain a novel class of implicature,

* We thank Judith Degen, Kai von Fintel, Danny Fox, Mike Frank, Michael Franke, Richard Futrell, Ted Gibson, Roni Katzir, Justine Kao, Dan Lassiter, Tim O'Donnell, Chris Potts, Greg Scontras, Raj Singh, Nathaniel Smith, and Benjamin Spector for comments and discussion. We also thank three anonymous reviewers and the editor, David Beaver, for extensive comments on an earlier draft. This work was supported by ONR grant N00014-13-1-0788, DARPA agreement number FA8750-14-2-0009, and a James S. McDonnell Foundation Scholar Award (NDG); by grant FG-BR2012-030 from the Alfred P. Sloan Foundation and a fellowship from the Center for Advanced Study in the Behavioral Sciences (RL); and by an NSF Graduate Research Fellowship and NSF Doctoral Dissertation Improvement Grant (LB).

non-convex disjunctive implicatures, which have several theoretically interesting properties. In particular, these implicatures can be preserved in downward-entailing contexts in the absence of accenting, a property which is predicted by lexical uncertainty, but which violates prior generalizations in the literature (Horn 1989, Fox & Spector 2018).

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 have historically met with less success than formalization in other linguistic domains such as phonology, syntax, or semantics. Even more successful models, such as Gazdar (1979)'s, do not attempt to model the sophisticated counterfactual reasoning which Grice describes in his discussions of particularized implicatures and flouting. 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 et al. 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 that are beginning to realize this potential by formalizing Gricean pragmatic reasoning using game theory or related decision-theoretic frameworks (Parikh 2000, Franke 2009, Jäger 2012, Franke & Jäger 2014, Rothschild 2013, Frank & Goodman 2012, Goodman & Stuhlmüller 2013, Degen, Franke & Jäger 2013, Benz, Jäger & Van Rooij 2005, Russell 2012). 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 (Frank & Goodman 2012, Goodman & Stuhlmüller 2013, Degen, Franke & Jäger 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 a 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 the multiple equilibrium problem for signaling games — a well-known problem in game theory. 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 our derivation of M-implicatures. By the term *inference* we refer to the use of data to estimate model parameters which are *a priori* unknown; by *semantic inference*, we are referring to the use of probabilistic inference to resolve the semantic content of utterances. Thus, the end result of pragmatic reasoning results from inferences about the meaning of words, not only about the speaker's intentions or beliefs. 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 exactly how their language's lexicon maps expressions to literal meanings. 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 novel 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 — that is, 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.

The phenomena discussed in this paper differ with respect to their novelty in the pragmatics literature and whether they can be explained under previous pragmatic accounts. Specificity implicatures (and their special case, scalar implicatures) are entirely standard in the game-theoretic pragmatics literature, and our account of these implicatures is essentially identical to previous proposals. M-implicatures have also been looked at extensively in this literature, but unlike specificity implicatures, there is no canonical explanation for them. We introduce a novel pragmatic principle, lexical uncertainty, to explain these implicatures. The final set of phenomena we consider, non-convex disjunctive implicatures, have not yet been considered in the pragmatics literature, and cannot be derived within previous game-theoretic accounts. We show how to derive these implicatures through a natural extension of the lexical uncertainty principle, thereby demonstrating an improvement in empirical coverage over these previous models. Non-convex disjunctive implicatures have further theoretical interest, because they can be derived in downward-entailing contexts, and therefore serve as counterexamples to previous generalizations in the literature. We will show that lexical uncertainty both explains the phenomena which motivated these generalizations, and provides an account of these counterexamples.

The models that we present are undoubtedly incomplete in many respects, and our goal is not to present a theory of pragmatics *per-se*. Rather, our goal is to present several new principles of pragmatic reasoning, and understand how these principles may be used to derive different classes of implicatures. These principles are, to the best of our knowledge, minimal sets of assumptions for deriving the phenomena considered in this paper within a probabilistic approach. These principles are therefore promising candidates for inclusion in more complete formal accounts of pragmatics. This is supported by an observation which will recur throughout the paper: the proposed principles are *conservative*, in the sense that extending simpler models with them preserves the major classes of implicatures derived by those simpler models. This supports the development of pragmatic theories in an incremental manner, and suggests that the ideas presented here may be incorporated into other accounts without disturbing the core predictions of those accounts.

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 (Lewis 1969, Stalnaker 1978, Clark 1996) — 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 $\mathcal{O}$ be the set of possible speaker observations, $\mathcal{W}$ the set of possible worlds, and $\mathcal{U}$ the set of possible utterances. Observations $o \in \mathcal{O}$ 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]

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 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:

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

To illustrate these definitions, consider a scenario in which the students in the class took a test, and the speaker has observed the test results for all of the students or none of them. In a simplified representation of this situation, there are two worlds,

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

corresponding to whether all of the students passed the test ($\forall$) or some but not all of them passed ($\exists\neg\forall$). There are three possible observations,

$$\mathcal{O} = \{\forall_o, \exists\neg\forall_o, \emptyset_o\},$$

corresponding to whether the speaker observed that all of the students passed ($\forall_o$), observed both that some passed and that not all passed ($\exists\neg\forall_o$), or did not make any relevant observations ($\emptyset_o$).[2] A possible joint probability distribution $P(o, w)$ is given by:

$$\begin{aligned} P(\forall_o, \forall) &= 0.25 \\ P(\exists\neg\forall_o, \exists\neg\forall) &= 0.25 \\ P(\emptyset_o, \forall) &= 0.25 \\ P(\emptyset_o, \exists\neg\forall) &= 0.25 \end{aligned}$$

There is probability 0.5 of all of the students passing the test, and given either state of the world ($\forall$ or $\exists\neg\forall$), the speaker has probability 0.5 of observing that state.

Continuing this example, we could set

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

with the intensions

$$\begin{aligned} [\![\text{some}]\!] &= \{\forall, \exists\neg\forall\} \\ [\![\text{all}]\!] &= \{\forall\} \end{aligned}$$

2 Note that given the limited space of worlds, not making any relevant observation is equivalent to observing that some passed, and not observing anything else.

The utterance “some" is therefore compatible with both worlds, while “all" is only compatible with $\forall$.

After hearing the utterance “all", the literal listener will exclude all worlds which are incompatible with with the meaning of the utterance. The only world compatible with this meaning is $\forall$, and therefore:

$$L_0(\forall_o, \forall | \text{all}) = 0.5$$
$$L_0(\emptyset_o, \forall | \text{all}) = 0.5$$

Only two observation-world pairs include the world $\forall$, so these are each assigned probability 0.5.

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 will say that the speaker has recursion level n if they reason about a listener with recursion level $n-1$; and that the listener has recursion level n if they reason about a speaker with recursion level n. This definition grounds out in the listener with recursion level 0, who has been defined in Equation (2). We begin with 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 communicate to the listener all of the information that the speaker has about the world, and low in utility insofar as they are costly to produce.

The expected utility of utterance u for a recursion-level n speaker who has observed o is defined as

$$U_n(u|o) = \mathbb{E}_{P(w|o)} \log L_{n-1}(o, w|u) - c(u). \tag{3}$$

The term $c(u)$ 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). The term $\mathbb{E}_{P(w|o)} \log L_{n-1}(o, w|u)$ is the negative EXPECTED SURPRISAL over observations and worlds given utterance u, and can be expanded as follows:

$$\mathbb{E}_{P(w|o)} \log L_{n-1}(o, w|u) = \textstyle\sum_w P(w|o) \log L_{n-1}(o, w|u) \tag{4}$$

The quantity $-\log L_{n-1}(o, w|u)$, the SURPRISAL, quantifies the residual information left about the observation o and world w after the listener L_{n-1} hears utterance u. The speaker wants to minimize the amount of infor-

mation that is left uncommunicated to the listener, and hence maximizes the negative surprisal. However, the speaker may not know what the true world is, and therefore may not know how much information is being left uncommunicated. The speaker uses the expected surprisal in Equation (4) to consider all of the worlds which are consistent with his observation, and average over the surprisal in each of these worlds. The speaker wants to minimize the expected amount of information that is left uncommunicated, while simultaneously minimizing the cost of their utterance.

In the first part of this paper, we assume that for each world $w \in \mathcal{W}$, there is a unique observation $o \in \mathcal{O}$ 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 expected surprisal reduces to the surprisal of the observation for the listener given the utterance.

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

$$U_n(u|o) = \log L_{n-1}(o|u) - c(u). \tag{5}$$

Section 3.1 shows how this formula can be used to derive scalar implicatures. 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

$$S_n(u|o) \propto e^{\lambda U_n(u|o)}, \tag{6}$$

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

$$L_n(o, w|u) \propto P(o, w)S_n(u|o). \tag{7}$$

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), (3), (6), and (7) constitute the heart of this basic model. Note the relationship between recursion levels of the speaker and listener in Equation (3): 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 (Grice 1989, Horn 1984, Gazdar 1979, Levinson 2000), our models of pragmatic reasoning will rely heavily on 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, that is, 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.[3]

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 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,

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

and two meanings,

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

where the intensions of the utterances are as usual:

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

3 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.

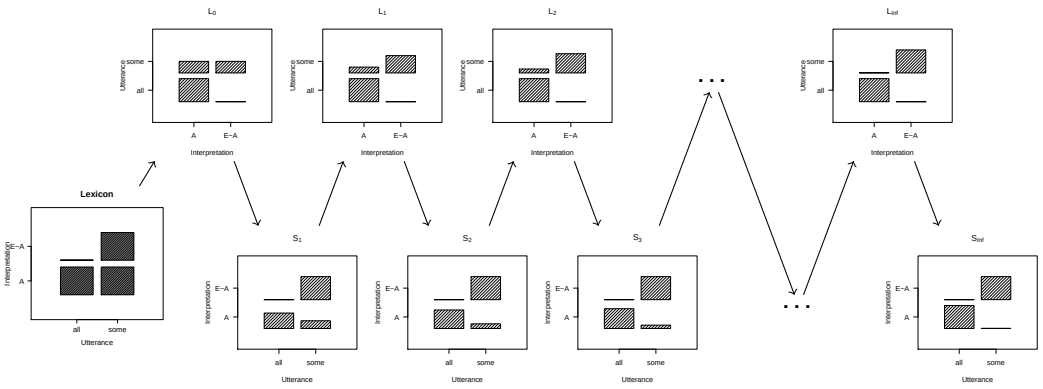


Figure 1 *Some* strengthening with $P(\forall) = \frac{1}{2}$, $P(\exists\neg\forall) = \frac{1}{2}$, $c(\text{all}) = c(\text{some}) = 0$, $\lambda = 1$. The lexicon panel indicates the truth value of utterances across each world. The listener panels indicate the conditional probabilities over worlds, given each utterance. The speaker panels indicate the conditional probabilities over utterances, given each world. Arrows are used to indicate dependence across the panels. The listener L_0 uses the lexicon in order to compute conditional probabilities given an utterance; the speaker S_1 uses the output of listener L_0 in order to compute utterance probabilities given each world; and so on. This figure, and several others in this form which appear later in the paper, are intended for readers who want to better understand the dynamics of the speaker-hearer recursion. The linguistic claims of the paper can be appreciated without relying on them.

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)–(7). This allows the baseline model to be expressed as

$$L_0(o|u,\mathcal{L}) \propto \mathcal{L}(u,o)P(o), \tag{8}$$

$$U_n(u|o) = \log L_{n-1}(o|u) - c(u), \tag{9}$$

$$S_n(u|o) \propto e^{\lambda U_n(u|o)}, \tag{10}$$

$$L_n(o|u) \propto P(o)S_n(u|o), \tag{11}$$

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 $\lambda = 1$.[4]

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; this represents the truth value of each utterance across the worlds. 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):

	Listener n	
all	$L_n(\forall\|\text{all})$	$L_n(\exists\neg\forall\|\text{all})$
some	$L_n(\forall\|\text{some})$	$L_n(\exists\neg\forall\|\text{some})$
	$\forall$	$\exists\neg\forall$

	Speaker n	
$\forall$	$S_n(\text{all}\|\forall)$	$S_n(\text{some}\|\forall)$
$\exists\neg\forall$	$S_n(\text{all}\|\exists\neg\forall)$	$S_n(\text{some}\|\exists\neg\forall)$
	all	some

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

4 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. For this example, and for others throughout the paper, we assessed robustness to changes in the model parameters by computing model predictions across a grid of parameter values.

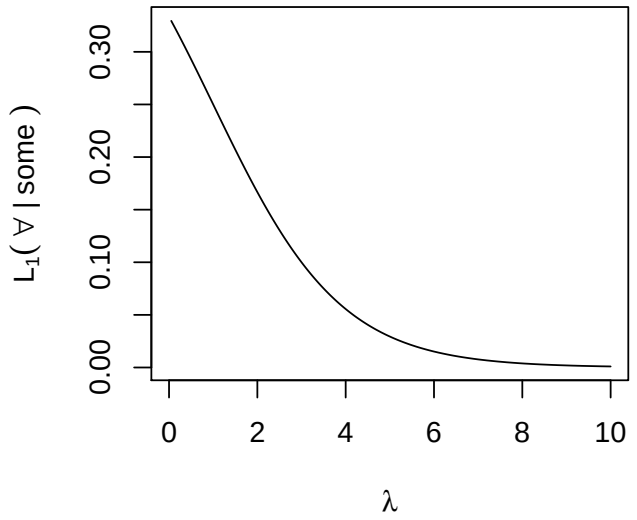


Figure 2 The degree of *some* strengthening as a function of the "greedy rationality" parameter λ, with $P(\forall) = \frac{1}{2}$, $P(\exists\neg\forall) = \frac{1}{2}$, $c(\text{all}) = c(\text{some}) = 0$

Equations (8)–(11):

L_{n-1}			U_n			S_n			L_n		
all	1	0	$\forall$	0	$\log p$	$\forall$	$\frac{1}{1+p}$	$\frac{p}{1+p}$	all	1	0
some	p	$1-p$	$\exists\neg\forall$	$-\infty$	$\log(1-p)$	$\exists\neg\forall$	0	1	some	$\frac{p}{2p+1}$	$\frac{1+p}{2p+1}$
	$\forall$	$\exists\neg\forall$		all	some		all	some		$\forall$	$\exists\neg\forall$

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 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, that is, 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:

(12) John can finish the homework.

(13) 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.[5]

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

5 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, that is, the expensive utterance was mapped to the least likely meaning, and so on.

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),

$$L_0(o, w|u, \mathcal{L}) = \frac{P(o, w)\mathcal{L}(u, w)}{\sum_{o', w'} P(o', w')\mathcal{L}(u, w')} \tag{14}$$

$$= \frac{P(o, w)\mathcal{L}(u', w)}{\sum_{o', w'} P(o', w')\mathcal{L}(u', w')} \tag{15}$$

$$= L_0(o, w|u', \mathcal{L}) \tag{16}$$

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 (3),

$$U_n(u|o) - c(u') = \mathbb{E}_{P(w|o)} \log L_{n-1}(o, w|u) - c(u) - c(u') \tag{17}$$

$$= \mathbb{E}_{P(w|o)} \log L_{n-1}(o, w|u') - c(u') - c(u) \tag{18}$$

$$= U_n(u'|o) - c(u) \tag{19}$$

It follows from equation (6) 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, the probability of the speaker using u and u' at any stage in the recursion 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 (7),

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

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

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

□

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 (Rabin 1990, Cho & Kreps 1987). 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

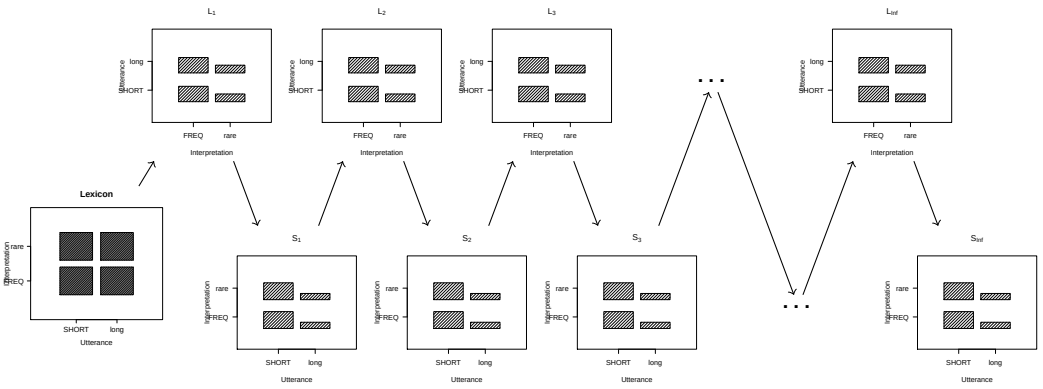
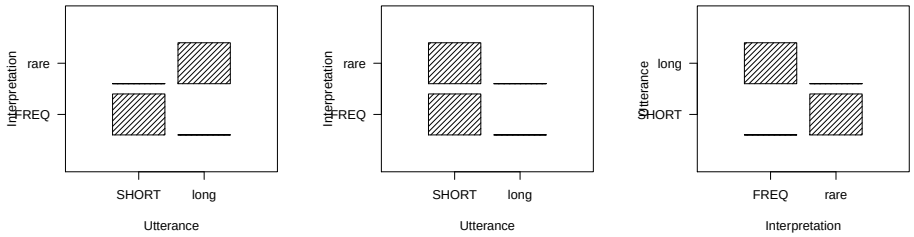


Figure 3 The failure of the basic model to derive M-implicature (illustrated here for $P(\text{FREQ}) = \frac{2}{3}$, $P(\text{rare}) = \frac{1}{3}$, $\lambda = 3$, $c(\text{SHORT}) = 1$, $c(\text{long}) = 2$). The listener panels illustrate that the interpretation of each utterance is constant across each recursion depth. The speaker panels illustrate that the utterance distributions are also constant across each recursion depth.

be played. The simplest equilibrium concept in game theory is the Nash equilibrium (Nash et al. 1950, Myerson 1991, Fudenberg & Tirole 1991). 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



(a) Separating Pareto-optimal

(b) Pooling 1

(c) Separating Pareto-suboptimal

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

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 4(a)). 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 4(c)). 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 4(b)). 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.[6] 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

6 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.

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 Van Rooy (2004) and De Jaegher (2008). 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 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 (2014) 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 (2014) 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 — that is, 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 (2014)'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 to the desired pragmatic effects.)

Introducing lexical uncertainty generalizes the previous model; the base listener L_0 remains unchanged from equation (2), that is, 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:

$$U_1(u|o, \mathcal{L}) = \mathbb{E}_{P(w|o)} \log L_0(o, w|u, \mathcal{L}) - c(u), \tag{27}$$

$$S_1(u|o, \mathcal{L}) \propto e^{\lambda U_1(u|o,\mathcal{L})}, \tag{28}$$

$$L_1(o, w|u) \propto P(o, w) \sum_{\mathcal{L}\in\Lambda} P(\mathcal{L})S_1(u|o, \mathcal{L}), \tag{29}$$

$$U_n(u|o) = \mathbb{E}_{P(w|o)} \log L_{n-1}(o, w|u) - c(u) \qquad \text{for } n > 1, \tag{30}$$

$$S_n(u|o) \propto e^{\lambda U_n(u|o)} \qquad \text{for } n > 1, \tag{31}$$

$$L_n(o, w|u) \propto P(o, w)S_n(u|o) \qquad \text{for } n > 1.^{7} \tag{32}$$

These equations differ from the baseline model in several respects. In Equations (27) and (28), the speaker S_1 is parameterized by a lexicon $\mathcal{L}$, which determines the speaker's beliefs about how their utterances will be interpreted. That is, this speaker believes that the listener L_0 will use this lexicon to interpret their utterances. The definition of the listener L_1 in Equation (29) is the most important difference between the current and baseline models. The listener L_1 in the baseline model (7) is certain about the speaker S_1's beliefs about the lexicon; for this listener, there is a single

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

(i) $U_n(u|o, w, \mathcal{L}) = \mathbb{E}_{P(o|w)} \log L_{n-1}(o, w|u, \mathcal{L}) - c(u).$

(ii) $S_n(u|o, w, \mathcal{L}) \propto e^{\lambda U_n(u|o,w,\mathcal{L})},$

(iii) $L_n(o, w|u, \mathcal{L}) \propto \sum_{\mathcal{L}'\in\Lambda} P(o, w)P(\mathcal{L}')S_n(u|o, w, \mathcal{L}'),$

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, based on the belief that they are easier to follow than those in Equations (i)–(iii).

lexicon which determines the speaker's beliefs about the literal meanings of utterances. In the current model, the listener L_1 (29) has uncertainty about which lexicon the speaker S_1 is using. For each possible lexicon $\mathcal{L} \in \Lambda$, the listener considers how the speaker would behave given this lexicon. To interpret an utterance, the listener first considers how likely the speaker would have been to choose this utterance given each lexicon, and then accounts for her uncertainty by marginalizing (taking a weighted average) over the lexica. The definitions of the higher-order speakers and listeners, in Equations (30)-(32), are the same as in the baseline model.

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

i. 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 $\mathcal{L}(u, w) > 0$.

ii. 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 ii 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 ii. 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 expected surprisal 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 largest 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 (that is, 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, $\exists w \ \mathcal{L}(u, w) > 0$. More informally, these conditions state that utterance meaning $\mathcal{L}(u, \cdot)$ is a valid refinement if it logically implies the semantic meaning $\mathcal{L}_S(u, \cdot)$, and if it is non-contradictory. 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:

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

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

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:

$$L_0(o|u, \mathcal{L}) \propto \mathcal{L}(u, o)P(o), \tag{34}$$

$$U_1(u|o, \mathcal{L}) = \log L_0(o|u, \mathcal{L}) - c(u), \tag{35}$$

$$S_1(u|o, \mathcal{L}) \propto e^{\lambda U_1(u|o,\mathcal{L})}, \tag{36}$$

$$L_1(o|u) \propto P(o) \sum_{\mathcal{L}\in\Lambda} P(\mathcal{L})S_1(u|o, \mathcal{L}), \tag{37}$$

$$U_n(u|o) = \log L_{n-1}(o|u) - c(u) \qquad \text{for } n > 1, \tag{38}$$

$$S_n(u|o) \propto e^{\lambda U_n(u|o)} \qquad \text{for } n > 1, \tag{39}$$

$$L_n(o|u) \propto P(o)S_n(u|o) \qquad \text{for } n > 1. \tag{40}$$

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

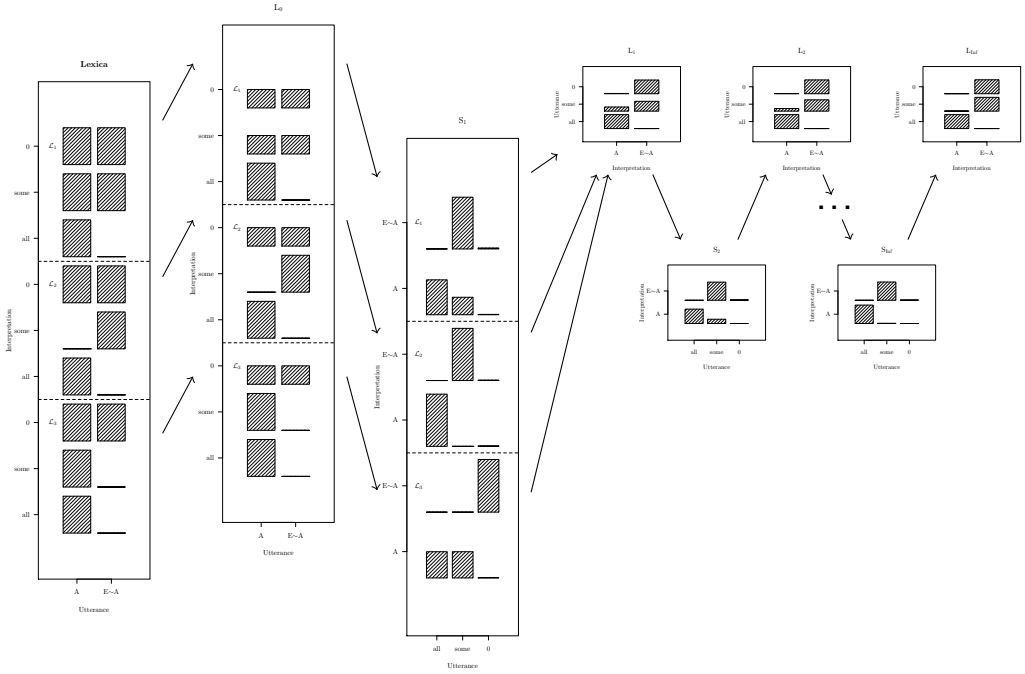


Figure 5 Specificity implicatures under lexical uncertainty, shown here with $P(\forall) = \frac{1}{2}$, $P(\exists\neg\forall) = \frac{1}{2}$, $c(\text{all}) = c(\text{some}) = 1$, $c(\emptyset) = 5$, $\lambda = 1$. The first column shows the three admissible lexica in this example. The second column illustrates the behavior of the listener L_0, with each row corresponding to a listener who is using a particular lexicon. The third column shows the speakers who correspond to these listeners. There is only a single distribution for listener L_1, as this listener computes utterance interpretations by averaging over the distributions for speaker S_1.

distribution over the speaker's intended meaning by marginalizing over the possible lexica that the speaker may have been using (Equation 37). As can be seen in the second column of the third panel of Figure 5, L_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 are still derived under lexical uncertainty.

It is important to note that the specificity implicature is not primarily driven by inferences about lexical content of "some." More precisely, the listener L_1 retains a high degree of uncertainty about the lexical content of "some" after hearing this utterance — much more uncertainty than they have about the *intended interpretation* of "some." As described above, if the listener hears "some," then there are multiple hypotheses about the speaker's communicative intent and their lexicon which will rationalize the choice of this utterance. Moreover, there are multiple lexica which are consistent with the speaker intending to communicate the world $\exists\neg\forall$ by this utterance. The speaker will use "some" to communicate $\exists\neg\forall$ if the lexical entry for "some" is $\exists\neg\forall$, and also if the entry is unenriched. As a result, even restricting to cases in which the listener L_1 has inferred that the speaker intends to communicate $\exists\neg\forall$, this listener will be uncertain about whether the lexical entry for "some" has been enriched. Pragmatic inference in this model thus *involves* resolution of the lexicon, but is not *identical* to lexical resolution.

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:

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

Each utterance has three possible enrichments — {FREQ, *rare*}, {FREQ}, and {*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 (37)), 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 in natural language but also

in one-shot rounds of simple signaling games (e.g. as observed in Bergen, Goodman & Levy (2012)).

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 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:

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

(42) Some of the students passed the test.

As discussed in Section 3, (42) has a specificity implicature that strengthens the literal meaning of "some" to an understood meaning of "some but not all". The implicatures of (41) differ crucially in two ways. First, as noted by Gazdar (1979; see also Chierchia, Fox & Spector 2012), (41) lacks the basic specificity implicature of (42). Second, (41) 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, Russell 2012, Meyer 2013, Fox 2014). This is for several reasons. First, the sentence in (41) 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 (that is, 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

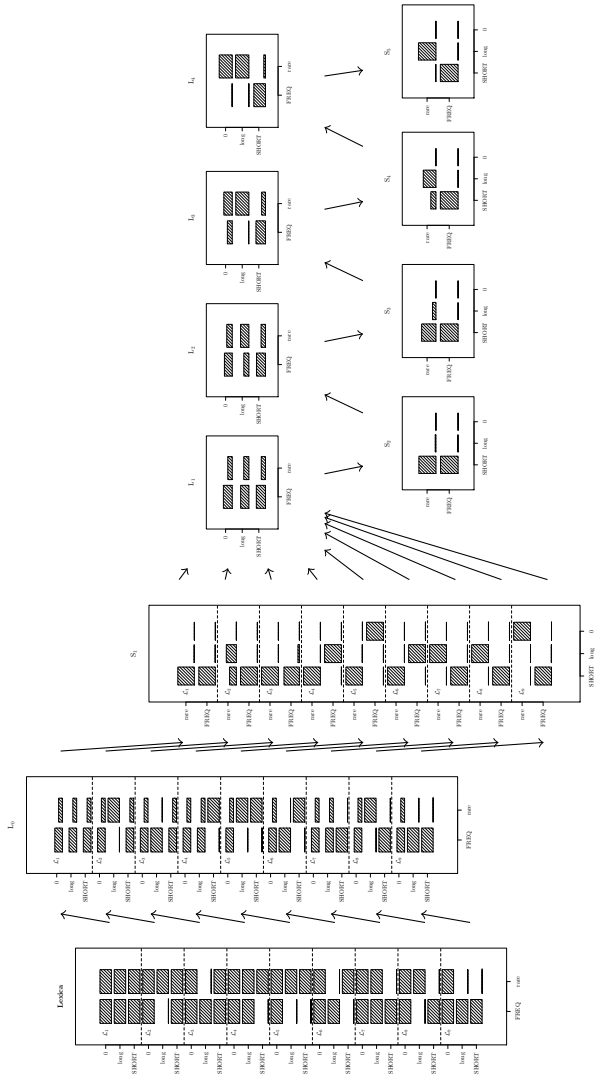


Figure 6 Deriving M-implicatures with $P(\text{FREQ}) = \frac{2}{3}$, $P(\text{rare}) = \frac{1}{3}$, $\lambda = 4$, $c(\text{SHORT}) = 1$, $c(\text{long}) = 2$, $c(\emptyset) = 5$. As in figure 5, the first column enumerates all of the admissible lexica, and the next two columns show the listener and speaker distributions corresponding to each of these lexica. The listener L_1 averages over the lexica in order to compute an interpretation for each utterance. Though small, there is already an asymmetry between the two utterances at listener L_1, with SHORT slightly more likely to be interpreted as FREQ.

distinct pragmatic interpretations for semantically equivalent expressions. To the best of our knowledge, there has been only one previous formal derivation of this class of implicatures, using an extension of the Iterated Best Response model (Jäger, Degen & Franke 2013).

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 (41) 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 (41) possess an ignorance implicature that corresponding sentences such as (42) do not should make the following empirically testable prediction. Consider a sentence pair like (41)–(42), differing only in TARGET QUANTIFIER "some or all" versus "some." For the "some or all" variant, comprehenders should be less likely to conclude that the speaker knows (a) exactly how many of the objects have the relevant property or (b) that *not all* of the objects have the relevant property. 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 more 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 $\lambda = 4$.[8] (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.

Although both can be derived with the same machinery, the ignorance implicature derived in this section is not just a repackaging of the M-implicatures derived in Section 4.5, but rather is a distinct phenomenon.

8 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 $P(w|o)$.

As shown in that section, lexical uncertainty will assign low-probability interpretations to complex utterances. In the current section, we have considered a scenario in which the interpretations receive uniform prior probability. In particular, the ignorant knowledge state is assigned the same probability as each state in which the speaker knows the true state of the world. Therefore, nothing in the assignment of prior probabilities breaks the symmetry between interpretations. We showed that lexical uncertainty nonetheless assigns the ignorant knowledge state as the interpretation of the complex utterance. This derivation of the ignorance implicature therefore exploits asymmetries between knowledgeable and ignorant knowledge states, rather than asymmetries between high-and-low probability states. Multiple forms of effective markedness emerge naturally from the lexical uncertainty model.

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

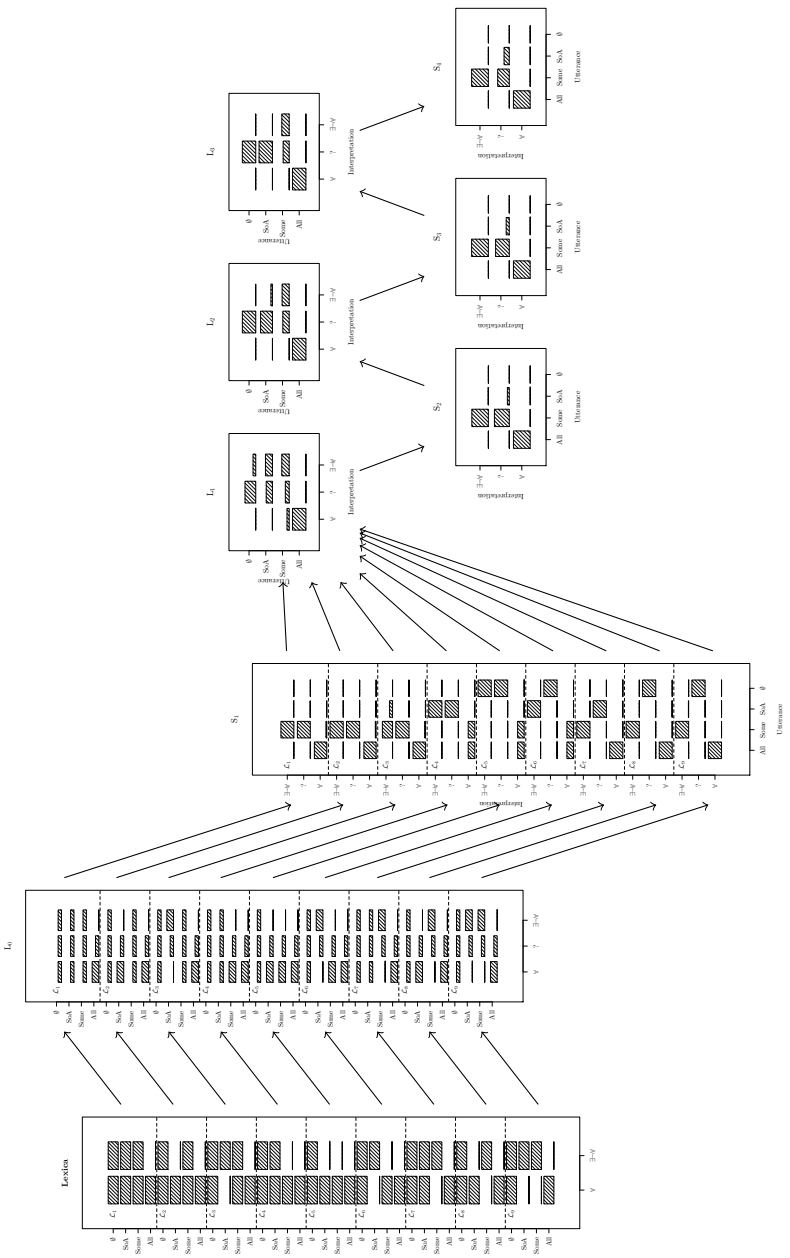


Figure 7 Deriving generalized markedness implicatures with $\lambda = 4$, uniform prior probabilities over observations, c("all")=0, c("some")=0, c("some or all")=1.

adopt the generalized notion of entailment from higher-order logics and then define a refinement of a lexical entry as another term of the same type that entails the original entry (Muskens 1995). The set of lexica Λ could then be derived, as before, as the set of all lexicons that can be derived from $\mathcal{L}_S$ 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 ii carries over: we may extend each lexicon with a trivial u_{null}. Restriction i 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[9]; we will initially 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

9 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-interpretable compositions (Goodman & Lassiter 2014).

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:

(43) 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.*]

(44) 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.*]

(45) Context: there are four people in a dance class, and at the beginning of each class, the 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 alternative 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*.[10] It is not at all clear on a standard neo-Gricean account how these pairs

10 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.

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*.[11] 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.

5.2 Compositional lexical uncertainty

In this section we further specify the compositional lexical uncertainty, as sketched out above, for the case of boolean atomic utterances composed

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.

11 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.

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:

$$[\![u]\!]_{\mathcal{L}}(w) = \mathcal{L}(u, w) \tag{46}$$

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":

$$[\![u_1 \text{or } u_2]\!]_{\mathcal{L}}(w) = \begin{cases} 1 & \text{if } [\![u_1]\!]_{\mathcal{L}}(w) = 1 \text{ or } [\![u_2]\!]_{\mathcal{L}}(w) = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{47}$$

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:

$$L_0(w, o|u, \mathcal{L}) \propto [\![u]\!]_{\mathcal{L}}(w) P(w, o) \tag{48}$$

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:

$$U_A = \{\textit{one}, \textit{two}, \textit{three}\} \qquad O = \quad \begin{array}{ccccc} & 1 & & 2 & & 3 \\ & | & \diagdown & & \diagdown & | \\ 1 \vee 2 & & & 1 \vee 3 & & 2 \vee 3 \\ & \diagdown & & | & \diagdown & \\ & & 1 \vee 2 \vee 3 & & & \end{array}$$

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

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

for the simple indicative case.

The set of possible lexica consists of all logically possible combinations of valid refinements (that is, 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):

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

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

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 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 $\{2,3\}$).

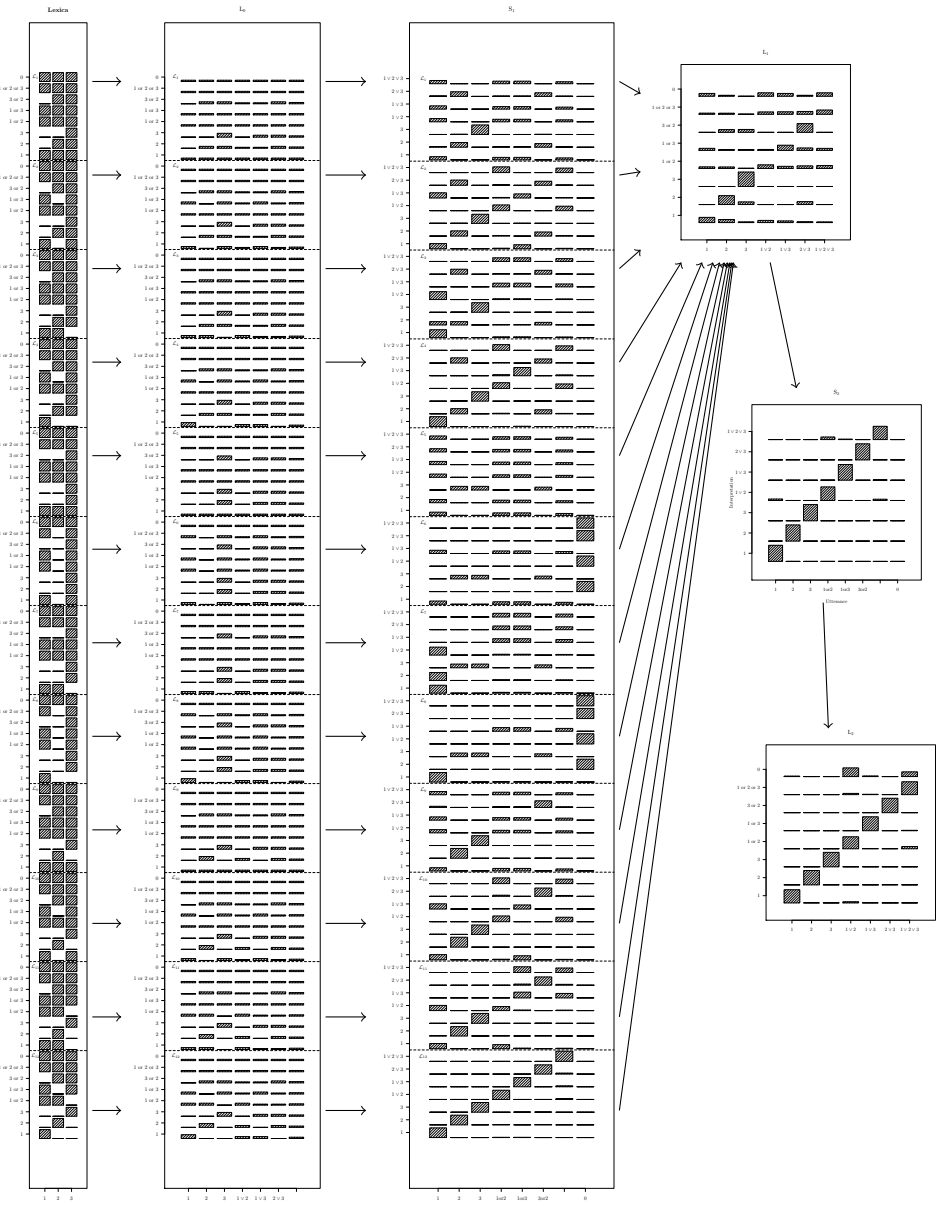


Figure 8 Non-convex disjunction, for uniform marginal distribution $P(\mathcal{O})$, uniform conditional distributions $P(W|O)$, cost per disjunct of 0.05, and $\lambda = 5$. Only lexica (and L_0 and S_1 distributions) in which the refined meaning of *one* contains the world state 1 are shown.

Consequently, the set of lexica in which *one or two* has $\{1, 3\}$ 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).[12]

5.5 Implicature cancellation

Does lexical uncertainty preserve standard properties which are associated with implicatures? In particular, does lexical uncertainty allow for the cancellation of implicatures? For example, consider the utterance “Some of the students passed the test, in fact they all did." Lexical uncertainty allows the literal meaning of “some" to be refined to mean $\{\exists\neg\forall\}$. If “some" is

12 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.

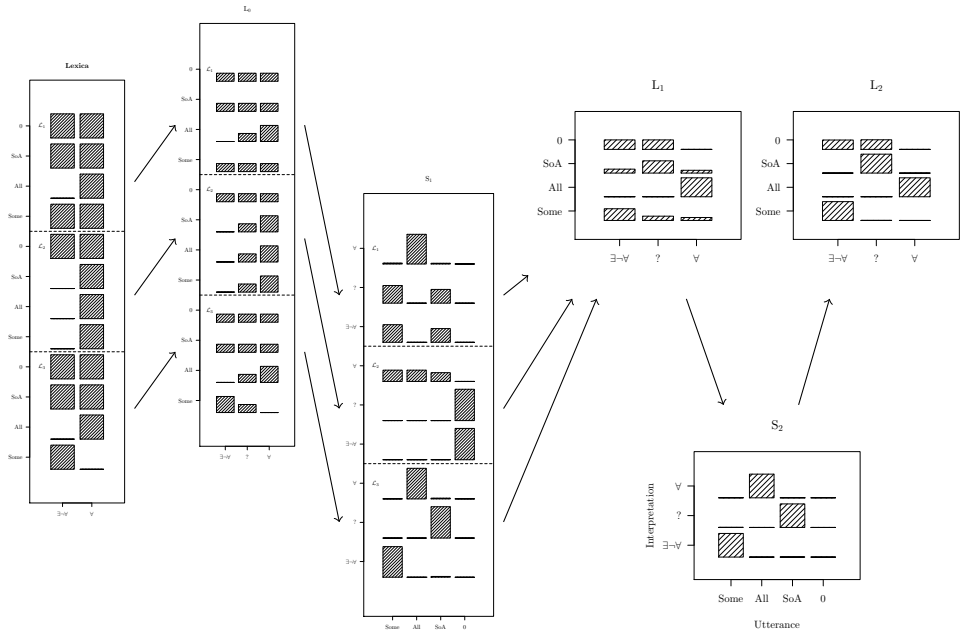


Figure 9 *Some or all* ignorance implicature under compositional lexical uncertainty.

refined in this manner, then the utterance will be contradictory, as it will assert that some but not all of the students passed the test, and that all did. Thus, it may appear that lexical uncertainty predicts that this utterance is contradictory — which would clearly a problem for our account.

Implicature cancellation is in fact possible under lexical uncertainty. As discussed in section 4.4, the derivation of the specificity implicature for “some" (or of other specificity implicatures) under the lexical uncertainty model is not primarily driven by the refinement of the lexical entry for “some." That is, the pragmatic listener has a high degree of certainty that the speaker intended to communicate $\{\exists\neg\forall\}$, but still considers it quite probable that the lexical entry for “some" is the unrefined $\{\exists\neg\forall, \forall\}$ (and that the narrow interpretation comes from the standard effect of alternatives). This property is retained in the compositional model: after hearing an utterance like “Some of the students passed the test," the listener L_1 will be uncertain about the lexical entry for “some."

Suppose that a listener then hears a cancellation utterance, such as the one above. In the compositional model, this utterance is treated as the

conjunction of two utterances: “some" and “all." If the listener L_1 only heard “some," they would be uncertain about whether its lexical entry was $\{\exists \neg \forall\}$ or $\{\exists \neg \forall, \forall\}$, or less probably $\{\forall\}$. However, given the conjunction of these two utterances, the listener is able to draw a stronger inference. If the lexical entry for “some" had been $\{\exists \neg \forall\}$, then the conjunction of “some" and “all" would have been contradictory, and the speaker would not have chosen the utterance in this case. The listener therefore will infer that the lexical entry of “some" was $\{\exists \neg \forall, \forall\}$. This lexical entry for “some" is consistent with $\forall$, and the literal content of “some and all" is $\forall$. The listener will therefore cancel the implicature, interpreting “some and all" as $\forall$.

5.6 Downward entailing contexts

We have shown how to derive a particular class of embedded implicatures using compositional lexical uncertainty. It is, moreover, possible to use the same machinery to straightforwardly derive many other standard embedded implicatures. However, certain constraints on these implicatures have been observed in the literature. We now consider whether lexical uncertainty can derive one such constraint: the observation that embedded implicatures generally do not occur in downward entailing contexts (Gazdar 1979, Horn 1989).

Consider the following example:

(49) John didn’t talk to Mary or Sue.

Without embedding under negation, as in (50) below, the disjunction would license a pragmatically strengthened, exclusive-or (XOR), meaning:

(50) John talked to Mary or Sue.

It has long been observed that when certain speaker knowledgeability assumptions are in the common ground, (50) indeed gives rise to this strengthened meaning of *John talked to either Mary or Sue, but not to both*. This a standard case of scalar implicature (through negation of the alternative generated by substitution of *and* for *or*) and falls out of all variants of our model, even without lexical uncertainty. If the disjunction were given this stronger XOR meaning within (49), then the resulting sentence meaning would be equivalent to *John talked to both Mary and Sue, or neither*. However, this appears to be a strongly dispreferred reading of (49), which seems to convey

that John did not talk to Mary, and that he did not talk to Sue. For grammatical approaches to embedded implicatures, which use an exhaustification operator to derive these implicatures, this observation has suggested that exhaustification operators cannot be applied in downward entailing contexts. Though exhaustification of the disjunction (through refinement of *or*) is not a possibility in our current formulation of lexical uncertainty, a nearly identical problem nonetheless arises for our approach, arising from the possibility of refinement of the lexical entries for the disjuncts. If we assign propositional representations to "Mary" and "Sue", denoted by M and S respectively, the propositional representation for (49) will be $\neg(M \vee S)$. Under one admissible refinement of these utterances, M will mean *John talked to Mary and not Sue* and S will mean *John talked to Sue and not Mary*. The expression $\neg(M \vee S)$ will in this case be equivalent to the unattested reading above: *John talked to both Mary and Sue, or neither*. Lexical uncertainty therefore predicts that the dispreferred XOR reading is available as a literal meaning of (49). It would be problematic for our theory if this reading were propagated through pragmatic reasoning, and assigned relatively high probability by the listener. We will show, however, that this is not the case: pragmatic reasoning under lexical uncertainty generally reduces the availability of the XOR interpretation, thus our theory predicts that this interpretation will be strongly dispreferred.

In order to demonstrate this, we will first present a formalization of Example (49) in our framework. We assume that there are four worlds:

$$\mathcal{W} = \{\{\}, \{\text{Mary}\}, \{\text{Sue}\}, \{\text{Mary, Sue}\}\}$$

Each world is specified by the set of people that John talked to. We assume all speaker epistemic states are possible: the set of observations $\mathcal{O}$ is maximal with respect to the worlds, that is, for every non-empty subset of $\mathcal{W}$, there is a corresponding observation that is consistent only with the worlds in that subset. We will use the term *knowledge state* to refer to the subset of worlds which are consistent with a particular observation. We assume that the prior distribution on observations, $P(o)$, is uniform, and that the conditionals on world given observation, $P(w|o)$, are each individually uniform (over the worlds which are compatible with that observation). (This implies that the marginal on worlds, $P(w)$, is uniform as well.) A refinement of an utterance is *compatible* with a knowledge state if the knowledge state is a subset of the refinement. We will also compare the *informativity* of utterances with respect to a given knowledge state o: of (refined) utterances u, u' both compatible

with o, u is more informative than u' with respect to o if u is compatible with fewer alternative knowledge states than u.

Utterances are assumed to be generated from the following grammar:

$$S \rightarrow C, \;\; S \rightarrow \neg C$$
$$C \rightarrow m \; L \; s, \;\; C \rightarrow m, \;\; C \rightarrow s$$
$$L \rightarrow \vee, \;\; L \rightarrow \wedge$$

This grammar derives two atomic utterances m and s, with

$$[\![m]\!] = \{\{\text{Mary}\}, \{\text{Mary, Sue}\}\}$$
$$[\![s]\!] = \{\{\text{Sue}\}, \{\text{Mary, Sue}\}\},$$

and more complex utterances which are formed through conjunction, disjunction, and negation. We assume that every utterance which is derived by the grammar is available as an alternative. We assume that an utterance has cost proportional to the total number of negations and disjunctions it contains, though our key qualitative predictions are invariant to precise utterance costs. Semantic refinement and composition are implemented in a manner nearly identical to the previous examples.[13]

For the models in this section and the next, we show the fixed-point interpretations to which the pragmatically sophisticated listener converges — we denote these as L_∞ — but the same qualitative behavior is apparent in the model before the fixed point is reached. Figure 10 shows the nine possible refined lexica for this model, and the ultimate predicted pragmatic interpretation of Example (49), as well as the interpretation of the other alternative utterances, when the cost per disjunct and the cost of negation are each 0.1, and the inverse-temperature parameter λ is 5.[14] The pragmatic listener

13 The only additional complication in this example is that certain combinations of refinements result in utterances with contradictory interpretations. For example, if the utterance $Mary$ is refined to mean that John talked to Mary but not Sue, and Sue is refined to mean that John talked to Sue but not Mary, then their conjunction will be contradictory. We adopt the solution briefly discussed in Section 5, and introduce a world state w_{err} which has positive probability if and only if all other world states have zero probability. The alternative utterance "John talked to Mary and Sue" maps to this world state in 5 of the 9 lexica, as shown in Figure 10. We assign zero prior probability to w_{err}, equivalent to the speaker and listener assuming joint communicative success (see also discussion in Footnote 9).

14 The figure shows one unintuitive prediction of the model: that $\neg(m \wedge s)$ will sometimes be interpreted as the fully ignorant knowledge state $\{\{\}, \{m\}, \{s\}, \{m, s\}\}$. The association of the fully ignorant knowledge state with this utterance occurs because $\neg(m \wedge s)$ has the least

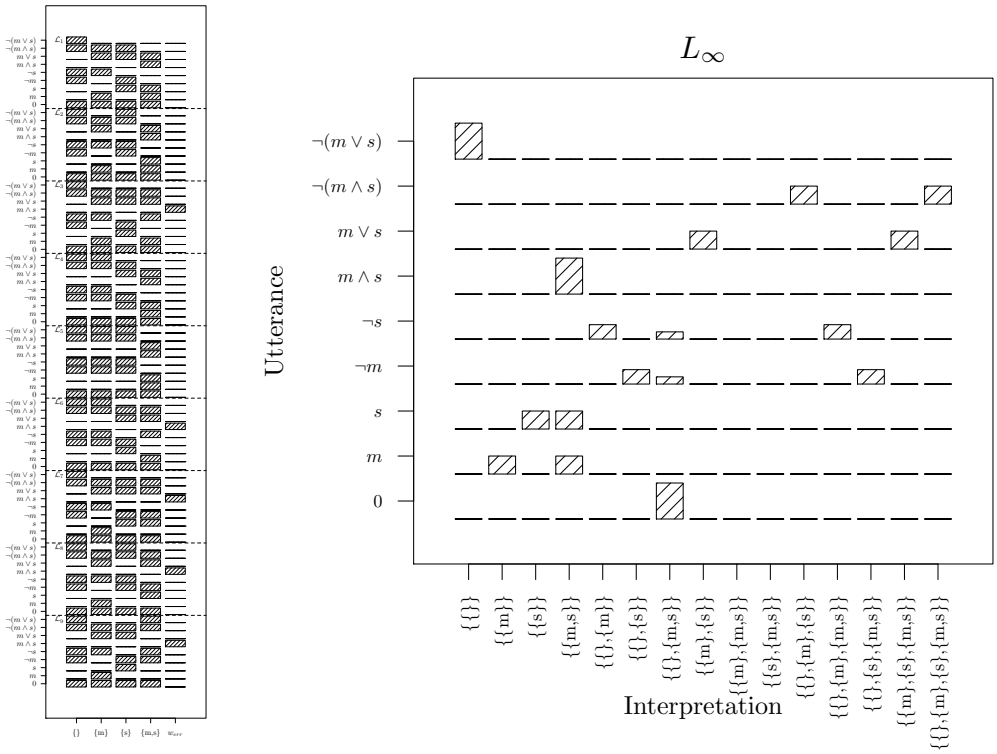


Figure 10 The set of lexica (left) in the negation/disjunction example (49), and the interpretation that the listener L_∞ assigns to each alternative utterance (right). The y-axis shows the 9 alternative utterances, while the x-axis shows the 4 possible worlds for the lexica, and the 15 possible knowledge states for the interpretation. The inverse-temperature $\lambda = 5$, the cost of each disjunct is set to 0.1, the cost of negation is 0.1, and the knowledge states receive uniform prior probability. Because there are a greater number of knowledge states than utterances, most utterances are not specialized to a single meaning. The critical prediction of the model — that utterance $\neg(m \vee s)$ will be interpreted as knowledge state $\{\{\}\}$ — is robust across all settings of the cost parameters which we have examined, ranging from 0 to 50.

interprets $\neg(m \vee s)$ as $\{\{\}\}$ with probability 1 — thus, Example (49) does not generate an embedded implicature in the model. To convey the key intuitions for this key behavior of the model, we will give a two-part explanation. First, we will explain why a speaker who wants to communicate this world will choose Example (49). Second, we will explain why the speaker would not choose $\neg(m \vee s)$ to convey any knowledge state other than $\{\{\}\}$. A pragmatically sophisticated listener using this knowledge to reason about the speaker will infer from $\neg(m \vee s)$ that the speaker intended knowledge state $\{\{\}\}$. Throughout this section and the next one, we will be focusing on the reasoning of the listener L_1, who interprets utterances by performing joint-inference over the speaker's knowledge state and lexicon. This listener's reasoning drives the effects which are discussed in this section; the reasoning of higher-order speakers and listeners mostly amplifies these effects.

To explain why the speaker who wants to communicate knowledge state $\{\{\}\}$ will choose Example (49), consider the speaker who wants to communicate this world. This speaker cannot use any utterance which entails that John talked to either Mary or Sue. For any pair of refinements to m and s, the only utterances which satisfy this requirement are those under the scope of negation, that is, those generated by the rule $S \rightarrow \neg C$. All four of these negated utterances — $\neg m$, $\neg s$, $\neg(m \wedge s)$, and $\neg(m \vee s)$ — are always compatible with knowledge state $\{\{\}\}$. Crucially, however, as can be seen in the left panel of Figure 10, under no refined lexicon is $\neg(m \vee s)$ less informative with respect to $\{\{\}\}$ than any of the other three negated utterances, and for each of these other three there are many refined lexica in which $\neg(m \vee s)$ is more informative: all but $\mathcal{L}_2$ and $\mathcal{L}_5$ for $\neg m$, all but $\mathcal{L}_4$ and $\mathcal{L}_5$ for $\neg s$, and all but $\mathcal{L}_5$ for $\neg(m \wedge s)$. The speaker who wants to communicate the world $\{\}$ will, under any lexicon, thus find $\neg(m \vee s)$ at least as good as any other utterance; and under most lexica, it will be better.

We will now explain why the speaker would be unlikely to use $\neg(m \vee s)$ to communicate any knowledge state other than $\{\{\}\}$. Note that there are a number of possible knowledge states besides $\{\{\}\}$ compatible with $\neg(m \vee s)$ under some refined lexicon, readable off of the lexica panel of Figure 10:

informative literal meaning among the non-null alternative utterances: depending on the lexicon, the literal meaning of this utterance is compatible with at least three, and often four, of the possible worlds. However, this association is sensitive to the cost of the null utterance, which is best at literally communicating the fully ignorant knowledge state. When the null utterance is sufficiently cheap, it will be used by the speaker in the fully ignorant knowledge state, and will therefore block the association of $\neg(m \wedge s)$ with this knowledge state.

$$\begin{array}{ll}
\{\{\},\{s\}\}, & \{\{\},\{m\},\{s\}\},\\
\{\{s\}\}, & \{\{m\},\{s\}\},\\
\{\{\},\{m\}\}, & \{\{\},\{m,s\}\},\\
\{\{m\}\}, & \{\{m,s\}\}.
\end{array}$$

We will give the explicit logic for why $\neg(m \vee s)$ is not the preferred utterance to express the critical XOR knowledge state {{}, {Mary, Sue}}; a similar logic applies to all the other knowledge states listed above. The first pragmatically sophisticated listener L_1 must reason about speaker S_1's behavior in the face of uncertainty about the lexicon that S_1 is using. But, as can be seen in the lexica panel of Figure 10, $\neg(m \vee s)$ is compatible with the XOR knowledge state {{}, {Mary, Sue}} in only one of the nine possible lexica ($\mathcal{L}_9$). The alternative utterance $\neg m$, in contrast, is compatible with the XOR knowledge state in three lexica, in two of which it is the most informative for expressing this knowledge state. The same is true for $\neg s$. The low prior probability of S_1 using a lexicon in which the utterance $\neg(m \vee s)$ is compatible with the XOR knowledge state immediately disadvantages this utterance for this state. This effect becomes stronger for more pragmatically sophisticated speakers, who never choose $\neg(m \vee s)$ to communicate {{}, {Mary, Sue}}, but rather $\neg m$ or $\neg s$.

The model implementation we have just described illustrates why embedded implicature for Example (49) is disfavored under compositional lexical uncertainty. Under all of the parameter settings we have explored, this disfavoring is strong enough to lead to complete unavailability of the locally strengthened interpretation after pragmatic inference. This does not mean, however, that overall model behavior is completely invariant to parameter settings. For example, a strongly skewed prior distribution over lexica which favors $\mathcal{L}_9$ could invalidate the second part of our logic as laid out above, and potentially lead to an XOR knowledge-state interpretation of $\neg(m \vee s)$. In this connection, we should note that we know of no empirical evidence showing that the XOR interpretation of Example (49) is categorically unavailable regardless of conversational context. We leave as an open empirical and modeling question whether there are conversational contexts in which listeners obtain strengthened XOR readings for utterances such as Example (49), and if there are, whether parameterizations of our model corresponding to features of such contexts lead to such readings.

In connection with this open question, it has previously been noted that embedded implicatures can be generated in downward-entailing environments if accenting is placed on the scalar term (Horn 1989):

(51) John didn't talk to Mary OR Sue.

Under one reading, this utterance is compatible with John having talked to both Mary and Sue. This suggests that the disjunction is being assigned an exclusive-or meaning, and therefore that an embedded implicature has been generated. We have not presented an account of how to treat accenting in our modeling framework. Whether accenting can be properly treated in this framework, and whether a proper treatment would derive the embedded implicature in Example (51), thus remain as additional open questions.

5.7 Exceptional downward entailing contexts

The examples discussed in Section 5.6 seem to provide evidence that, in the absence of accenting, embedded implicatures are not preserved in downward-entailing contexts. Indeed, there are several theoretical proposals which have been developed to account for this generalization (Chierchia, Fox & Spector 2012, Fox & Spector 2018). Chierchia, Fox & Spector (2012), who derive embedded implicatures using exhaustivity operators, propose the following condition: an exhaustivity operator cannot be inserted into a sentence if it results in an interpretation which is logically weaker than what the sentence would receive in its absence. This straightforwardly accounts for the unavailability of the embedded implicature in Example (49), as the reading associated with the implicature is logically weaker than the attested reading. Fox & Spector (2018) propose an extension of this condition, which maintains the implication that exhaustification is excluded when it results in a logically weaker reading.

We will present evidence, however, that these generalizations do not hold in all circumstances. In particular, it is possible to construct counterexamples using the non-convex disjunctive implicatures identified in Section 5.1. Consider the following examples:

(52) Context: A and B are visiting a resort. B has very particular preferences about the temperature of the springs at the resort: he will bathe in them if they are between 85-95 ℉ (30-35 ℃), or between 105-115 ℉ (40-45 ℃), as he finds the lower temperatures relaxing and the higher

temperatures invigorating. A knows about B's preferences, and has checked the water temperature for him.

A: The water isn't warm or scalding. [Understood meaning: the water is below 85 ℉ or between 95-105 ℉.]

(53) Context: A and B are scientists who study cancer in mice. They are discussing a tumor that one mouse has developed. If it is above 1 mm in size, then it cannot be removed safely. If it is between 0.1-1 mm, then it can be surgically removed, and the mouse can be saved; if it is between 0.01-0.1 mm, then it is too small to be surgically removed, but may still be harmful to the animal; and if it is less than 0.01 mm, then it is so small that it will not harm the animal. These facts about mouse tumors are common knowledge among A and B, and A has gotten some information about the tumor size.

A: The tumor isn't small or microscopic. [Understood meaning: the tumor is larger than 1 mm or between 0.01-0.1 mm.]

In example (52), the speaker embeds the Hurford-violating disjunction "warm or scalding" under negation. As discussed in Section 5.1, this disjunction ordinarily generates the non-convex implicature *warm but not hot, or scalding*. The current example demonstrates that this implicature can be preserved under negation, as the utterance is interpreted as the negation of *warm but not hot, or scalding*, that is, *cool or hot but not scalding*. This provides evidence both that embedded implicatures can be generated in downward-entailing contexts, and that Hurford-violating disjunctions can be felicitous in such contexts (though see Fox & Spector 2018, which proposes a related example, in which a Hurford-violating disjunction is felicitously embedded in a downward-entailing context, *without* its embedded implicature being preserved).

We have already shown that lexical uncertainty can explain the lack of embedded implicatures in Example (49). We will now show that it can simultaneously explain the embedded implicature generated in Example (52). It is the distinctive structure of the sets of worlds and alternatives in Example (52) which lead to this implicature. The conditions which prevented an embedded implicature from being generated in Example (49) are absent in this example.

In order to formalize this example, we assume that the set of worlds $\mathcal{W} = \{1, 2, 3, 4\}$, where higher numbers correspond to higher temperatures. World 1 corresponds to water below 85 ℉, world 2 to water between 85-95 ℉,

world 3 to water between 95-105 ˚F, and world 4 to water between 105-115 ˚F. The speaker's possible knowledge states are again the non-empty members of $2^{\mathcal{W}}$, the powerset of $\mathcal{W}$. The speaker's knowledge states receive uniform prior probability. There are three atomic utterances, with the following intensions: $[\![w]\!] = \{2, 3, 4\}$ ("warm"), $[\![h]\!] = \{3, 4\}$ ("hot"), $[\![s]\!] = \{4\}$ ("scalding"). The full set of utterances consists of the atomic utterances, disjunctions of arbitrary subsets of the atomic utterances, and the negations of the previous two types of utterances.[15] The utterance in Example (52) is represented by $\neg(w \vee s)$ in this formalization.

Figure 11 shows the predicted pragmatic interpretation of each alternative utterance, when, as in Figure 10, per-disjunct and per-negation costs are 0.1 and the inverse-temperature constant λ is 5. This behavior is qualitatively different than that seen for Example (49) as discussed in Section 5.6: here, $\neg(w \vee s)$ conveys knowledge state $\{1, 3\}$ (that is, the water is either warm or scalding, but not hot) with probability 1, corresponding to an interpretation with local scalar strengthening of w ("warm").

Why the difference in the behavior of Examples (49) and (52) in our compositional lexical uncertainty model? In our discussion the utterance in Example (49), we outlined the following logic: for an utterance u and knowledge state o, (i) is it the case that o is better expressed by u than by any other utterance? And (ii) will any alternative knowledge state $o' \neq o$ preferentially be expressed by u? If we can answer (i) in the affirmative and (ii) in the negative, then the listener should preferentially interpret u as conveying o. In the case of Example (49), we were able to answer (i) in the affirmative and (ii) in the negative for the knowledge state corresponding to no local strengthening, or $\{\{\}\}$. In Example (52), in contrast, it is the observation $\{1, 3\}$ that allows us to answer (i) in the affirmative and (ii) in the negative, corresponding to an interpretation with local strengthening of w. As with the discussion in Section 5.6, a key component of the reasoning lies in considering the number of different refined lexica in which various literal interpretations are available. Figure 13 in Appendix C depicts the 21 refined lexica for this problem, though for discursive simplicity we will not refer directly to specific lexica in this section.

15 We do not include conjunctions of the utterances, as in most cases, they are literally equivalent to the stronger conjunct. In particular, because $[\![s]\!] = \{4\}$, any conjunction which includes the utterance s will be either contradictory or literally equivalent to s. Including the conjunctions as alternatives does not, however, substantially change the predictions discussed here.

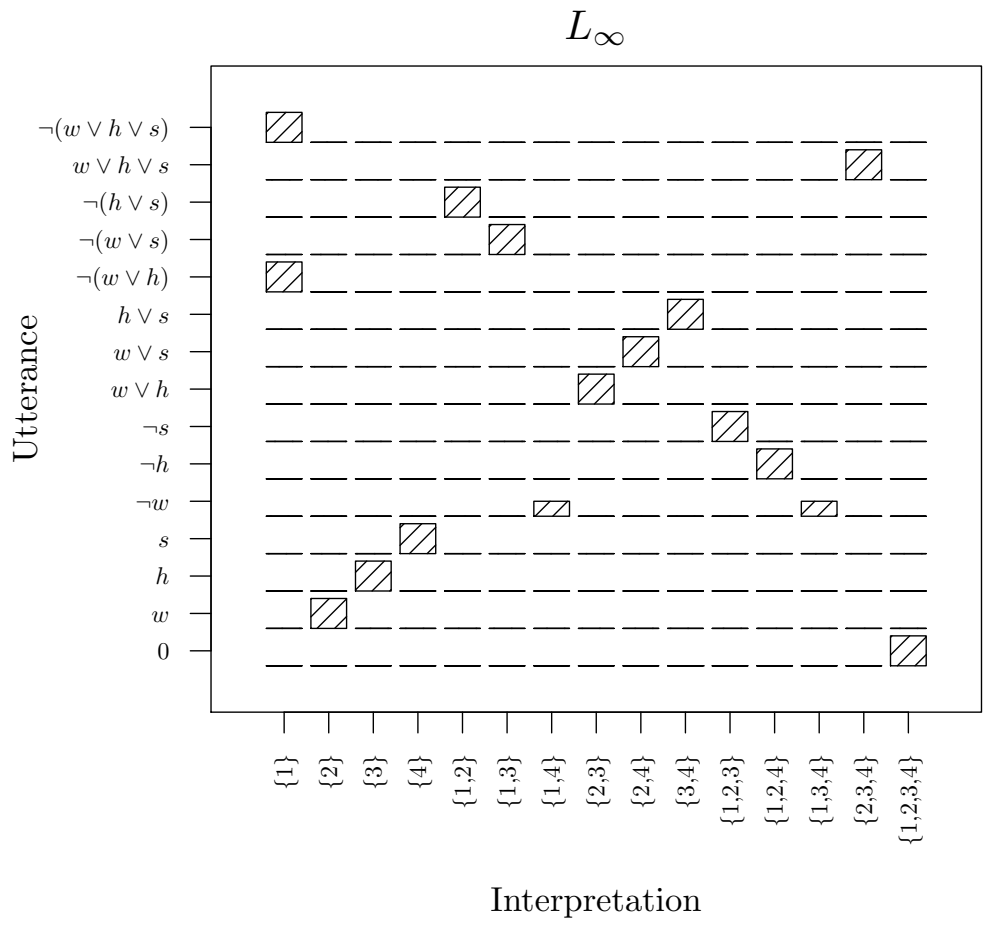


Figure 11 The interpretation the listener L_∞ assigns to each alternative utterance, in our formalization of Example (52). The y-axis shows the 15 alternative utterances, and the x-axis shows the 15 possible knowledge states. The cost of each disjunct is 0.1, the cost of negation is 0.1, the inverse-temperature λ is set to 5, and all knowledge states receive uniform prior probability.

We elucidate this logic first by explaining part (i): why the speaker with knowledge state $\{1,3\}$ will use Example (52). The utterance $\neg(w \vee s)$ is always literally compatible with world 1, due to the fact that refinements of the atomic utterances w and s must always be monotonic enrichments. The utterance is literally compatible with knowledge state $\{1,3\}$ in nine of the twenty-one possible lexica: six in which w is strengthened to $\{2\}$ or $\{2,4\}$, and three in which w is strengthened to $\{4\}$. In the first six, the literal (post-refinement) meaning of $\neg(w \vee s)$ conveys $\{1,3\}$ as informatively as possible. No other utterance has this degree of compatibility and informativity with respect to this knowledge state. $\neg s$ is always literally compatible with the knowledge state, but is never maximally informative, and in a number of lexica is less informative than an alternative. $\neg w$ is literally compatible with $\{1,3\}$ in the same nine lexica as $\neg(w \vee s)$, but is maximally informative in only three of them, and thus less informative overall. Similarly, no other alternative is as informative as $\neg(w \vee s)$.

We now turn to (ii): why no knowledge state other than $\{1,3\}$ will be preferentially expressed with $\neg(w \vee s)$. We lay out the explicit logic for the two most crucial alternative knowledge states to consider, namely $\{1\}$ and $\{3\}$, but similar logic applies to other knowledge states compatible with some refined literal meaning of $\neg(w \vee s)$. Knowledge state $\{1\}$, in which the speaker knows that the water is cool, would be the only knowledge state compatible with $\neg(w \vee s)$ without an embedded implicature. But the speaker will not choose $\neg(w \vee s)$ given knowledge state $\{1\}$, because there are alternative utterances which direct the listener more informatively toward this knowledge state. Of the seven possible refinements of w, only two (appearing in six of the 21 possible refined lexica) result in a literal interpretation of $\neg(w \vee s)$ as world 1; the other five yield weaker and thus less informative literal meanings. The utterance $\neg(w \vee h)$, in contrast, has literal interpretation $\{1\}$ in eight of the 21 refined lexica. The speaker in knowledge state $\{1\}$ is therefore more likely to choose utterance $\neg(w \vee h)$ than $\neg(w \vee s)$. We now consider knowledge state $\{3\}$, in which the speaker knows that the water is hot but not scalding. The utterance $\neg(w \vee s)$ is compatible with $\{3\}$ in only eight of the 21 lexica — those in which world 3 is not in the refinement of w. Moreover, even in these eight lexica, $\neg(w \vee s)$ is not maximally informative as to this knowledge state, as the utterance is always compatible with world 1. In contrast, the utterance h ("hot") is compatible with $\{3\}$ in 14 lexica, and uniquely picks out this knowledge state in seven of them. The speaker in knowledge state $\{3\}$ will therefore prefer utterance h over utterance $\neg(w \vee s)$.

The logic above characterizes why interpretations corresponding to embedded implicatures occur in our model under negation for non-convex disjunctions, even under circumstances where they do not appear for the more ordinary disjunctions explored in Section 5.6. Unlike the case of Section 5.6, however, where qualitative model behavior was invariant to utterance costs and the inverse-temperature λ, embedded implicature for non-convex disjunctions is sensitive to these model parameters. In particular, once the per-(disjunct/negation) cost rises beyond a threshold, the interpretation matrix depicted in Figure 11 crucially changes: the utterance $\neg(w \vee s)$ loses its embedded implicature and takes interpretation $\{1\}$, and the interpretation of utterance $\neg s$ is split 50/50 between $\{1,3\}$ and $\{1,2,3\}$. The precise cost threshold at which this change occurs depends on the inverse-temperature parameter λ: the cost threshold is about 0.35 for $\lambda = 5$, the value used in Figure 11; the threshold is higher for lower values of λ.

Intuitively, this change of interpretation arises because the additional cost of the disjunction eliminates the pragmatic blocking effects of disjunctive utterances. As noted above, for the speaker who wants to communicate knowledge state $\{1\}$, the most informative utterance is $\neg(w \vee h)$. When the cost of disjunctive utterances is sufficiently low, the speaker will always want to choose this utterance, and as a result, the utterance induces a blocking effect: if the speaker did not choose utterance $\neg(w \vee h)$, then that fact indicates that they are not in knowledge state $\{1\}$. When the cost of disjunctive utterances increases, this blocking effect disappears. If the speaker did not choose utterance $\neg(w \vee h)$, there are now two explanations for this fact: either the speaker is not in knowledge state $\{1\}$, or the speaker *is* in this knowledge state, and they decided that the utterance is too expensive. When the speaker uses utterance $\neg(w \vee s)$, the interpretation $\{1\}$ is no longer blocked by utterance $\neg(w \vee h)$. The utterance $\neg(w \vee s)$ *does* provide information about knowledge state $\{1\}$, and as a result this knowledge state is a reasonable interpretation of the utterance. The additional cost of disjunction also means that $\neg(w \vee s)$ does not exhibit any pragmatic blocking effects. In particular, though $\neg(w \vee s)$ is interpreted as knowledge state $\{1\}$, it does not block the interpretation of other utterances as this knowledge state. The simpler utterance $\neg w$ assigns $\frac{1}{3}$ of its probability mass to this knowledge state.

The above discussion has illustrated that compositional lexical uncertainty model has the expressive resources to account for the embedded implicature in Example (52), though it does not predict this implicature in

all cases. We briefly note that the implicature in Example (52) does in fact appear to be quite fragile. In the absence of any background context, the utterance does not appear to generate an implicature, or at least not as strongly (though further experimental evidence is needed to evaluate this intuition). The modeling setup presented in this section did not include any of this background context. In particular, the prior distribution was uniform over the speaker's possible knowledge states. Given different background information, such as low prior probability to knowledge states which are irrelevant in the example, the model predicts the implicature more robustly across utterance cost differentials. We leave for future work more comprehensive discussion, modeling, and empirical testing of these issues.

To sum up, then: contrary to previous claims in the literature, we have argued that embedded implicatures can be generated in downward-entailing contexts, and Hurford-violating disjunctions may be permissible in these contexts. We have further shown that lexical uncertainty can be used to generate these implicatures. Yet there are other downward-entailing contexts in which embedded implicatures cannot be generated. Lexical uncertainty accounts for this heterogeneity: the structure of the (un-refined) lexical denotations and conversational context determines whether an embedded implicature will be generated in a downward-entailing environment.

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 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 explored embedded implicatures, focusing on 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.

We further showed that lexical uncertainty predicts heterogeneous effects within downward-entailing contexts: while many embedded implicatures are canceled within downward-entailing contexts, some with particular scale structure can survive.

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 Interpretations of lexical uncertainty

Lexical uncertainty posits that the literal interpretation of a word-string is not fully fixed prior to its use in a conversational context. On its own, this claim is clearly not distinctive or new. Any string containing a lexically ambiguous word will also have its literal meaning left unfixed prior to use. A natural question is therefore whether lexical uncertainty is equivalent to positing an especially pervasive type of lexical ambiguity. Under such an interpretation, the different possible refinements of a word's semantic content correspond to different senses of that word in the lexicon. For example, in the simplified setting considered above, the word "some" would have three senses in the lexicon, corresponding to its three admissible refinements, $\{\forall, \exists\neg\forall\}, \{\forall\}, \{\exists\neg\forall\}$. The conditions on admissible refinements from Section 4.3 would be used to determine which senses of a word to include in the lexicon; however, this representation seems un-parsimonious given that all of these senses are systematically related.

If lexical uncertainty amounts to a variety of lexical ambiguity, then ordinary lexical ambiguities could be resolved according to the principles considered in this paper. This is a non-trivial commitment; it makes several predictions about ambiguity resolution, that must be empirically evaluated. We sketch these predictions in case they are useful in provoking future work.

The most straightforward implementation of word-sense ambiguity resolution in the lexical uncertainty framework treats the union of senses as the underlying meaning, and thus allows refinement to any one of the senses. For example, the word "bank" could be refined to the financial sense or the river sense (or left un-refined as a place that is both river and financial, though this is presumably ruled out by world knowledge). The model predicts that the listener will prefer word senses which are compatible with high probability states of the world. For instance, if Bob is a banker, then "Bob went to the bank" will tend to be interpreted in its financial sense. However, the model

also predicts that the listener will generally prefer disambiguating to an informative word sense rather than an uninformative word sense. For example, consider a situation in which it is common knowledge that the speaker went to a financial bank, but there is uncertainty about whether the speaker also went to a river bank. The lexical uncertainty account predicts that the listener will disambiguate “bank" to its river sense, because the alternative sense would have provided little information in the context. Word-sense disambiguation will be an interaction of these pressures (as well, potentially as linguistic factors, such as frequency of use for each sense, not treated in the models of this paper).

If these predictions are correct it would suggest a unification of disambiguation and the richer implicature phenomena considered in this paper. On the one hand implicature would be seen as a combination of disambiguation and the Gricean effect of alternatives. On the other hand ambiguity would be given a formal description and extended in scope to include additional cases not normally noticed. Alternatively, the above predictions may be incorrect, in which case ambiguity and lexical uncertainty are different flavors of uncertainty in language understanding.

6.2 The granularity of lexical refinement

A key aspect of our proposal is that lexical meanings admit refinements, and that a pragmatic listener reasons about which refinement is in use in a given context. We have remained mute on the appropriate granularity of context up to now. That is, at what level of temporal, or discourse, detail are refinements individuated? At one extreme, there is a single true refinement of each word that the pragmatic listener spends her whole life trying to pin down. In this case lexical refinement amounts to lexical *learning*. At the other extreme, a separate refinement is entertained for every use of every word: lexical uncertainty is *token-level*. In between these two extremes, refinements could vary from sentence to sentence or from conversation to conversation — realizing a form of semantic *adaptation*.[16] The examples we have considered in this paper have been restricted to have only one token of each word, and hence are insensitive to the granularity of refinement.

16 Another possibility is that refinements can vary at multiple timescales, being perhaps more conservative at longer timescales. A hierarchical Bayesian prior could capture this notion, effectively unifying short timescale adaptation and long timescale learning.

It will be important to explore the granularity of refinement in future work. It is, however, surprisingly difficult to find phenomena which clearly distinguish between possibilities. Lexical refinement only has an indirect effect on interpretation, that is, it is possible to derive pragmatic interpretations which are stronger than the inferred lexical refinement, and also possible for strengthened lexical refinements to have little effect on interpretation. Consider, as an example, the sentence “Some of the children laughed, and some of the adults laughed, in fact they all did." On first glance it seems that this sentence must be interpreted by assigning a different refinement to each token of “some”: the first restricted, *some but not all*, the second unrestricted, *at least some*. However, the correct interpretation can be achieved even if both tokens are given the unrestricted meaning: in the lexical uncertainty model alternative utterances still affect interpretation, and alternatives are still entertained independently for each token.

As discussed in section 4.4, when the listener L_1 hears the utterance “Some of the children laughed" on its own, they will derive a specificity implicature, and will infer that not all of the children laughed. Although the specificity implicature associated with this utterance is quite strong, the inference about the lexical content of “some" is weak: the listener is uncertain whether “some" has been refined to mean *some but not all* or *at least some*, that is, both of these refinements are compatible with the implicature. Now consider again the sentence “Some of the children laughed, and some of the adults laughed, possibly all of them." Suppose both instances of “some" are assigned the refinement *at least some*. The listener will still draw the specificity implicature (that is, not all) for “Some of the children laughed,” because the specificity implicature is not dependent on a strengthened literal meaning for “some." At the same time, the refinement for “some" is literally compatible with all of the adults having laughed. In conjunction with the speaker’s claim that all of the adults did laugh, the utterance “Some of the adults laughed" will be interpreted strongly: they all did. This example therefore demonstrates that it is not possible to simply read lexical refinements off of available pragmatic interpretations; relatedly that restrictions on lexical refinements do not straightforwardly translate to restrictions on pragmatic interpretations.

6.3 The flexibility of lexical refinement

In section 4.3, we suggested a procedure for building the set of lexica, Λ, from an underlying propositional meaning. This procedure allowed nearly any refinement of an utterance's propositional content. Is it possible that this flexibility would have unpleasant consequences in more complex models? For instance, if we simply relax the simplifying assumptions we made on spaces of worlds in sections 4.4 and 4.5, additional refined meanings become available. If the world contains an *is raining?* feature, then "some of the students passed the exam" could be refined to mean "some of the students passed the exam and it is raining." Is it the case that such spurious refinements will lead to incorrect interpretations? While it is beyond the scope of this paper to address the issue conclusively, we see three reasons to be optimistic that the lexical uncertainty approach is relatively robust to spurious refinements (or can be made so). These three reasons rely on symmetry along irrelevant dimensions, ignorance of irrelevant dimensions, and the regularizing effect of a question under discussion.

Recall the example of specificity implicatures used in section 4.4. In that example, there were only two worlds, $\forall$ and $\exists\neg\forall$ (representing the number of students who passed the test). The worlds in this example were individuated in a maximally coarse-grained manner: we collapsed all worlds which did not differ with respect to whether all of the students passed the test. This had the consequence of restricting the set of possible refinements of the items in the lexicon. In particular, the utterance "all" had only one possible refinement, the set $\{\forall\}$, and the utterance "some" had only three possible refinements, corresponding to the three non-empty subsets of the set of worlds. If the worlds had been individuated in a more fine-grained manner, then there would have been a greater number of possible lexical refinements. For example, suppose that the worlds were individuated along two dimensions: whether all of the students passed the test and whether it is raining outside. This would produce four worlds: $\forall \wedge R$,$\forall \wedge \neg R$, $\exists\neg\forall \wedge R$, and $\exists\neg\forall \wedge \neg R$. In this scenario, there are considerably more possible refinements of the lexicon. For "all," there are now three refinements: $\{\forall \wedge R$, $\forall \wedge \neg R\}$,$\{\forall \wedge R\}$, and $\{\forall \wedge \neg R\}$. For "some," there are 15 refinements, corresponding to the nonempty subsets of the four worlds. Many of these refinements carry propositional content which would never be conveyed by the lexical items in an actual conversation. For example, if "all" is refined to $\{\forall \wedge R\}$, then "all" will imply that it is raining outside. Such spurious implications will be

possible whenever worlds are allowed to vary along dimensions which are orthogonal to the utterances' semantic content. What effect do these spurious refinements have on the model's overall predictions?

In many simple cases, there are symmetries among the different spurious refinements, which have the consequence that the pragmatic listener gains no information about irrelevant dimensions. Suppose in the example above that the speaker is fully knowledgeable about both the number of students who passed the test, and whether it is raining. If "all" is refined to $\{\forall \wedge R\}$, then it will communicate that it is raining outside. Similarly, if it is refined to $\{\forall \wedge \neg R\}$, then it will communicate that it is not raining. The listener L_1, however, is uncertain about which refinement was being used by the speaker. As a result, the listener will not gain any information about whether it is raining, and their pragmatic interpretation of "all" (and similarly "some") will remain the same.

Suppose, on the other hand, that the speaker knows how many of the students passed the test, but does not have any information about the weather. In that case, if the refinement of the utterance communicates any information about the weather, then the speaker will not use it — to use the utterance in this case would communicate something that the speaker does not know, which is something that the speaker never does. As a result, when the listener hears the utterance, they will infer that it communicated no information about the weather, and thus their pragmatic interpretation of the utterance will be the same as before.

This reasoning can be generalized, to provide one sufficient condition under which spurious refinements will not communicate any spurious information to the listener. In appendix D, we show that the listener will not gain any information about the dimensions along which the speaker is ignorant despite the availability of spurious refinements of lexical items in these dimensions. The intuition for this result is the same as in the example above: the speaker will never choose an utterance which communicates information about the unknown dimensions, and therefore the listener will infer that their perceived utterance only provides information about the known dimensions.

Yet it seems that ignorance is a stronger condition than is needed to avoid spurious entailments — *irrelevance* also seems to be enough. Consider, for example, a speaker who is fully knowledgeable about how many students passed the test and whether it is raining, but who only cares about answering the question *Did all of the students pass the test?* In this case, however the alternative utterances are refined with respect the weather, they should

have no effect on the speaker's choice of utterance. As a result, the listener will not gain any information about the weather from the speaker's choice of utterance. This reasoning can be formalized by combining the question under discussion (QUD) extension to RSA proposed in (Kao et al. 2014, Kao, Bergen & Goodman 2014, Lassiter & Goodman 2015) with the lexical uncertainty extension. Using much the same argument, it is possible to show that, if the QUD is common knowledge between the speaker and listener, then the speaker will never communicate information about dimensions which are collapsed (irrelevant) under the QUD.

We have suggested three situations (symmetry, ignorance, irrelevance) under which spurious entailments will be resisted by the lexical uncertainty model despite the availability of spurious refinements. A different theoretical option is to embrace much stronger restrictions on the possible refinements: only refine along the "direction" of semantic content, or in other lexically specified ways. Indeed, the *free variable* formulation of lexical uncertainty lends itself naturally to this approach.

This alternative formalization uses *semantic free variables* as a locus of uncertainty.[17] 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 the semantic free variable technique can be seen as an instance of the general principle of lexical uncertainty (by viewing the lexica as the set of groundings of the free variables), it diverges from the flexible refinement procedure we have used for the results in this paper. In the latter, all refinements of an initial meaning are considered; the meanings that are generated by filling in free variables may be a very restricted subset of these. An impor-

17 Lexical uncertainty, as first proposed in Bergen, Goodman & Levy (2012), assumes that the lexicon fully specifies the semantic content of each lexical item. The formulation in terms of semantic free variables was proposed in Lassiter & Goodman (2013).

tant 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.

6.4 Psychological implications of the models

The models presented in this paper make predictions about how utterances will be interpreted, in different types of communicative contexts. They are intended as theories of pragmatic competence, or, in the terminology of Marr (1982), as computational-level theories of pragmatic knowledge. This is a surprisingly subtle claim, due to how the models were defined. We defined the models in a procedural manner, first specifying how the listener L_0 interprets utterances, then how the speaker S_1 chooses utterances given L_0, etc. The output of the model is defined to be the output of this process, either stopping at a particular recursion depth or iterating until a fixed point. Given the procedural nature of this definition, it may at first appear that we are making a claim about the algorithmic implementation of pragmatic reasoning: when people perform pragmatic reasoning, they reason in this recursive manner. Though this is possibly true, this is not our intended claim. The recursive procedure associated with the models is primarily a device for defining these models, that is, for defining the function from utterances to their pragmatic interpretations. We do not currently know of alternative ways of defining the relevant functions from utterances to interpretations, but such alternative definitions may exist, and may not involve a similar recursive formulation. We have not presented any evidence in this paper which would adjudicate between different methods for defining these functions; to the extent that we are presenting a theory of pragmatic competence, no such evidence *could* exist.

In particular, the psychological process which implements our competence model is constrained only loosely. It is possible that people reason recursively online, in the way suggested by our model's recursion. It is more likely, given the computational demands of such a process, that people at least *cache* their previous inferences, and possibly use a different procedure

altogether. Understanding the process of pragmatic inference consistent with our approach will first require a better understanding of the space of algorithms that can be used for inference — itself a hard theoretical question.

A more empirically tractable question concerns the correct parameters of our model. There are a number of latent parameters in our pragmatics models, which jointly determine pragmatic interpretation: utterance costs, recursion depth, the inverse-temperature λ, etc. In general, estimating the pragmatics model that best fits an individual's or population's inferences requires jointly estimating these parameters. Some existing empirical work suggests that the recursion depth is low (about L_1) and λ is moderate but greater that one (Goodman & Stuhlmüller 2013, Degen, Franke & Jäger 2013, Franke & Degen 2015). More work is needed to pin this down completely, especially for the complex implicatures dealt with here. A further complication for addressing this psychological question is that people may not reason to the same depth across scenarios. It is possible, for example, that they reason to higher depths when this is easy to do so (e.g. when they have encountered similar scenarios in the past), and lower depths when it is hard to do so. This issue complicates the interpretation of experimental results showing, for example, that people only compute to a particular depth in games which were introduced to them in the laboratory (Ho, Camerer & Weigelt 1998, Camerer, Ho & Chong 2004).

A related question concerns how precisely people calibrate their reasoning to particular conversational contexts. Speakers may, for example, optimize their utterances for an "average" listener, or may optimize for the particular conversation that they are in. Though it may appear that optimizing for an average interlocutor will decrease the cost of pragmatic inference, this is in fact not so clear. For the speaker to be able to communicate effectively under such a scheme, the listener would need to know what the speaker considers an average listener. Otherwise, the speaker may systematically communicate meanings which they did not intend to communicate. There are a large number of possibilities here; it is not clear if any will reduce computational complexity relative to case-by-case optimization, and which will produce successful protocols for communication. The experimental findings in this area mirror this complexity: there is no consensus on the extent to which speakers in these experiments optimize for the local conversational context (Nadig & Sedivy 2002, Hanna, Tanenhaus & Trueswell 2003, Keysar, Lin & Barr 2003).

6.5 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.

We follow many previous authors in using cost to derive certain pragmatic inferences. Grice's Maxim of Manner provides the following conversational norm (Grice 1989): "Be brief (avoid unnecessary prolixity)." Grice illustrated the use of this maxim with the following example: "Miss X produced a series of sounds that corresponded closely with the score of 'Home sweet home.'" The speaker uses a needlessly complex expression to describe Miss X's act of singing, and therefore violates the maxim; the listener infers from this violation that the speaker did not want to convey an ordinary sequence of events, and that Miss X's singing must have been abnormal in some way. In general, when utterances are (apparently) too complex, they will lead to violations of this maxim, and will trigger an implicature suggesting that something unusual happened. Horn uses his division of pragmatic labor in order to describe the effect of utterance complexity on interpretation, and to derive phenomena which are similar to those that Grice treated with the Maxim of Manner (Horn 1984). Horn explicitly proposes that brief or simple utterances will tend to be assigned common meanings, while longer or more complex utterances will tend to be assigned uncommon meanings. He thus predicts that M-implicatures, as we have called them in this paper, represent a systematic pattern of interpretation in natural languages. For Horn, the division of pragmatic labor is itself derived from competition between two more basic pragmatic principles, the Q-Principle and the R-Principle. Levinson's M-Principle is closely related to the division of pragmatic labor (Levinson 2000): it states that marked expressions will be used to describe abnormal situations, while unmarked expressions will be used to describe normal situations.

Our approach shares several features with these previous accounts. We represent speakers as preferring utterances with lower cost; *ceteris paribus*, an increase in an utterance's cost leads to a decrease in the speaker's utility.

As in the case of the Maxim of Manner and Horn's R Principle, this induces a listener expectation that the speaker will not use utterances which are overly complex. When the listener hears an (apparently) overly complex utterance, they try to rationalize this choice of utterance, that is, they try to find a meaning which would have made the use of this utterance rational. The reasoning here is quite similar to the reasoning which follows a violation of the Maxim of Manner under Grice's account, or a violation of the R Principle under Horn's account. In both of those cases, the speaker's apparent violation triggers a search for alternative interpretations of the utterance, which would render the speaker's use of the utterance appropriate. Under Grice's and Horn's accounts, as well as the current one, the interpretive effects of cost are derived from the listener trying to avoid attributing irrationality to the speaker.

Though cost plays an important role in prior accounts as well as the current one, there is no consensus on the proper operationalization of cost. That is, there is no consensus about which precise features of an utterance determine its costliness to the speaker. 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 affect 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.

6.6 Embedded implicatures

Cohen (1971) was the first to identify embedded implicatures as a challenge for Gricean theories of pragmatics. In that work, it was observed that implicatures associated with conjunction ordering can be preserved under embedding. Consider the following examples, slightly modified by Carston (1988):

(54) The old king died of a heart attack and a republic was declared.

(55) If the old king died of a heart attack and a republic was declared Sam will be happy, but if a republic was declared and the king died of a heart attack Sam will be unhappy.

Example (54) conveys information about the temporal ordering of events: the king first died of a heart attack, and a republic was declared subsequently. Grice (1989) proposed that the conjunction in this example has a classical semantics, and analyzed the temporal ordering inference as an implicature, which is derived from his orderliness maxim. Example (55) presents a problem for this analysis. In this example, the sentence in (54) has been embedded in a conditional, but it still conveys ordering information: Sam will be happy if the king first died of a heart attack and then a republic was declared, but not if a republic was first declared and then the king died of a heart attack. In other words, it appears that the conditional applies both to the semantic content of the embedded phrase, and to the ordering implicature that this phrase generates when it is asserted. Like the other embedded implicatures discussed in this paper, this provides a *prima facie* problem for the Gricean account. It is not straightforward to extend the Gricean derivation of the orderliness implicature of Example (55) to this case.

Relevance theorists such as Sperber & Wilson (1986) and Carston (1988) (though see Levinson 2000 for related suggestions) propose an alternative

account of semantic and pragmatic content, which can explain cases like Example (55). Under this account, the semantic content of the embedded phrase in Example (55) is underdetermined, that is, its propositional content is unfixed prior to its use in the sentence. Following the assertion of Example (55), the listener uses pragmatic inference to determine the semantic content of the antecedent of the conditional. During pragmatic inference, the listener concludes that this embedded phrase conveys information about the ordering of events — that it conveys the proposition that the king died first, and then a republic was declared — and this ordering information is included as part of the semantic content of the embedded phrase. The utterance (55) therefore carries an *explicature*, so that the antecedent of the conditional literally conveys a certain ordering of events, and the conditional as a whole literally conveys that Sam will be happy if this ordering of events holds (and unhappy if the reverse ordering of events holds). The theory posits that a single pragmatic principle — the expectation that utterances will be relevant — drives inferences about both the pragmatic content of utterances and their semantic content.

Like relevance theory, our account posits that the literal interpretation of utterances is underdetermined, and that pragmatic inferences can intrude on literal interpretation. Also like relevance theory, our account states that the same pragmatic principles which are used to determine pragmatic inferences also are used to fill in the literal content of utterances. At this high level of abstraction, the accounts primarily differ in the pragmatic principles which they use to derive these inferences. While relevance theory uses its relevance principle, our account uses a game-theoretic formalization of Gricean reasoning principles, which have been amended with the lexical uncertainty principle in order to explain the flexibility of literal interpretation. The accounts also clearly differ in the degree of their formalization. The question of whether and how relevance theory can be formalized is, to the best of our knowledge, an outstanding question in the field.

The particular class of embedded implicatures that we have focused on were not identified in the relevance theory literature, but rather by those supporting grammatical accounts of implicature computation (Chierchia, Fox & Spector 2012). We have focused on implicatures arising from Hurford-violating disjunctions 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 (Chier-

chia, Fox & Spector 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 (Sauerland 2004, Russell 2006). 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.

Equally important, the class of models we have introduced here captures a wide variety of M-implicatures (such as doing something unusual to "get the car started"), which are not addressed by theories based on exhaustification. Correctly predicting embedded scalar implicatures while unifying them with this broader set of implicatures represents an important expansion of empirical coverage.

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

Benz, Anton, Gerhard Jäger & Robert Van Rooij (eds.). 2005. *Game theory and pragmatics*. New York, NY: Palgrave Macmillan. https://doi.org/10.1057/9780230285897.

Bergen, Leon, Noah D Goodman & Roger Levy. 2012. That's what she (could have) said: how alternative utterances affect language use. In Naomi Miyake, David Peebles & Richard P. Cooper (eds.), *Proceedings of the 34th annual conference of the Cognitive Science Society*, 120–125. Austin, TX: Cognitive Science Society.

Camerer, Colin F, Teck-Hua Ho & Juin-Kuan Chong. 2004. A cognitive hierarchy model of games. *The Quarterly Journal of Economics* 119(3). 861–898. https://doi.org/10.2139/ssrn.2676402.

Carston, Robyn. 1988. Implicature, explicature, and truth-theoretic semantics. In Ruth Kempson (ed.), *Mental representations: the interface between language and reality*, 155–181. Cambridge University Press.

Chierchia, Gennaro. 2006. Broaden your views: implicatures of domain widening and the "logicality" of language. *Linguistic Inquiry* 37(4). 535–590. https://doi.org/10.1162/ling.2006.37.4.535.

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, 2297–2331. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110253382.2297.

Cho, In-Koo & David M Kreps. 1987. Signaling games and stable equilibria. *The Quarterly Journal of Economics* 102(2). 179–221. https://doi.org/10.2307/1885060.

Clark, Herbert H. 1996. *Using language*. Cambridge University Press. https://doi.org/10.1017/cbo9780511620539.

Cohen, L Jonathan. 1971. Some remarks on Grice's views about the logical particles of natural language. In *Pragmatics of natural languages*, 50–68. Springer. https://doi.org/10.1007/978-94-017-2020-5_5.

De Jaegher, Kris. 2008. The evolution of Horn's rule. *Journal of Economic Methodology* 15(3). 275–284. https://doi.org/10.1080/13501780802321400.

Degen, Judith, Michael Franke & Gerhard Jäger. 2013. Cost-based pragmatic inference about referential expressions. In Markus Knauff, Michael Pauen, Natalie 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. https://doi.org/10.1057/9780230210752_4.

Fox, Danny. 2014. Cancelling the Maxim of Quantity: another challenge for a Gricean theory of scalar implicatures. *Semantics and Pragmatics* 7(5). 1–20. https://doi.org/10.3765/sp.7.5.

Fox, Danny & Roni Katzir. 2011. On the characterization of alternatives. *Natural Language Semantics* 19(1). 87–107. https://doi.org/10.1007/s11050-010-9065-3.

Fox, Danny & Benjamin Spector. 2018. Economy and embedded exhaustification. *Natural Language Semantics* 26. 1–50. https://doi.org/10.1007/s11050-017-9139-6.

Frank, Michael C & Noah D Goodman. 2012. Predicting pragmatic reasoning in language games. *Science* 336(6084). 998–998. https://doi.org/10.1126/science.1218633.

Franke, Michael. 2009. *Signal to act: game theory in pragmatics*. Universiteit van Amsterdam dissertation.

Franke, Michael & Judith Degen. 2015. Reasoning in reference games: individual- vs. population-level probabilistic modeling. *manuscript, Tübingen & Stanford*.

Franke, Michael & Gerhard Jäger. 2014. Pragmatic back-and-forth reasoning. In Salvatore Pistoia Reda (ed.), *Pragmatics, semantics and the case of*

scalar implicatures, 170–200. Palgrave Macmillan. https://doi.org/10.1057/9781137333285.0011.

Fudenberg, Drew & Jean Tirole. 1991. *Game theory*. MIT Press.

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

Goodman, Noah D & Daniel Lassiter. 2014. Probabilistic semantics and pragmatics: uncertainty in language and thought. *Handbook of Contemporary Semantic Theory, Second Edition*. https://doi.org/10.1002/9781118882139.ch21.

Goodman, Noah D & Andreas Stuhlmüller. 2013. Knowledge and implicature: modeling language understanding as social cognition. *Topics in Cognitive Science* 5(1). 173–184. https://doi.org/10.1111/tops.12007.

Grice, H Paul. 1989. *Studies in the ways of words*. Harvard University Press.

Hanna, Joy E, Michael K Tanenhaus & John C Trueswell. 2003. The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language* 49(1). 43–61. https://doi.org/10.1016/s0749-596x(03)00022-6.

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. https://doi.org/10.1287/mnsc.14.3.159.

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

Ho, Teck-Hua, Colin Camerer & Keith Weigelt. 1998. Iterated dominance and iterated best response in experimental "p-beauty contests". *American Economic Review* 88(4). 947–969.

Horn, Laurence. 1984. Toward a new taxonomy for pragmatic inference: q-based and R-based implicature. In Deborah Schiffrin (ed.), *Meaning, form, and use in context*, 163–192. Washington, DC: Georgetown University Press.

Horn, Laurence. 1989. *A natural history of negation*. University of Chicago Press.

Hurford, James R. 1974. Exclusive or inclusive disjunction. *Foundations of Language* 11(3). 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*, 2487–2425. De Gruyter Mouton. https://doi.org/10.1515/9783110253382.2487.

Jäger, Gerhard, Judith Degen & Michael Franke. 2013. The iterated best response model of game theoretic pragmatics and its relatives. In Social

Dynamics Conference, UC Irvine. http://www.sfs.uni-tuebingen.de/~gjaeger/slides/slidesIrvine.pdf.

Kao, Justine T, Leon Bergen & Noah D Goodman. 2014. Formalizing the pragmatics of metaphor understanding. In Paul Bello, Marcello Guarini, Marjorie McShane & Brian Scassellati (eds.), *Proceedings of the 36th annual meeting of the Cognitive Science Society*, 719–724. Austin, TX: Cognitive Science Society.

Kao, Justine T, Jean Y Wu, Leon Bergen & Noah D Goodman. 2014. Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences* 111(33). 12002–12007. https://doi.org/10.1073/pnas.1407479111.

Keysar, Boaz, Shuhong Lin & Dale J Barr. 2003. Limits on theory of mind use in adults. *Cognition* 89(1). 25–41. https://doi.org/10.1016/s0010-0277(03)00064-7.

Lassiter, Daniel & Noah D Goodman. 2013. Context, scale structure, and statistics in the interpretation of positive-form adjectives. In *Proceedings of Semantics and Linguistic Theory*, vol. 23, 587–610. https://doi.org/10.3765/salt.v0i0.2658.

Lassiter, Daniel & Noah D Goodman. 2015. Adjectival vagueness in a Bayesian model of interpretation. *Synthese*. 1–36. https://doi.org/10.1007/s11229-015-0786-1.

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

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

Lewis, David. 1970. General semantics. *Synthese* 22(1). 18–67. https://doi.org/10.1007/bf00413598.

Lewis, Richard L. & Shravan Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science* 29(3). 1–45. https://doi.org/10.1207/s15516709cog0000_25.

Marr, David. 1982. *Vision: a computational approach*. San Francisco, CA: Freeman & Co.

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. https://doi.org/10.1007/978-94-010-2506-5_10.

Muskens, Reinhard. 1995. *Meaning and partiality*. Stanford, CA: CSLI.

Myerson, Roger B. 1991. *Game theory: analysis of conflict*. Harvard University Press.

Nadig, Aparna S & Julie C Sedivy. 2002. Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychological Science* 13(4). 329–336. https://doi.org/10.1111/j.0956-7976.2002.00460.x.

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. https://doi.org/10.1023/A:1005513919392.

Rabin, Matthew. 1990. Communication between rational agents. *Journal of Economic Theory* 51(1). 144–170. https://doi.org/10.1016/0022-0531(90)90055-0.

Rothschild, Daniel. 2013. Game theory and scalar implicatures. *Philosophical Perspectives* 27(1). 438–478. https://doi.org/10.1111/phpe.12024.

Russell, Benjamin. 2006. Against grammatical computation of scalar implicatures. *Journal of Semantics* 23(4). 361–382. https://doi.org/10.1093/jos/ffl008.

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. https://doi.org/10.1023/b:ling.0000023378.71748.db.

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

Sperber, Dan & Deirdre Wilson. 1986. *Relevance: communication and cognition*. Harvard University Press.

Stalnaker, Robert. 1978. Assertion. In Peter Cole (ed.), *Syntax and semantics 9: pragmatics*, 315–332. Academic Press.

Sutton, Richard S & Andrew G Barto. 1998. *Reinforcement learning: an introduction*. Cambridge, MA: MIT 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. https://doi.org/10.1126/science.1192788.

Van Rooy, Robert. 2004. Signalling games select Horn strategies. *Linguistics and Philosophy* 27(4). 493–527. https://doi.org/10.1023/b:ling.0000024403.88733.3f.

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:

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

(57) Some of the students passed the test.

The experiment evaluated two claims about the interpretation of example (56). The first claim is that while example (57) implicates that not all of the students passed the test, example (56) 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.

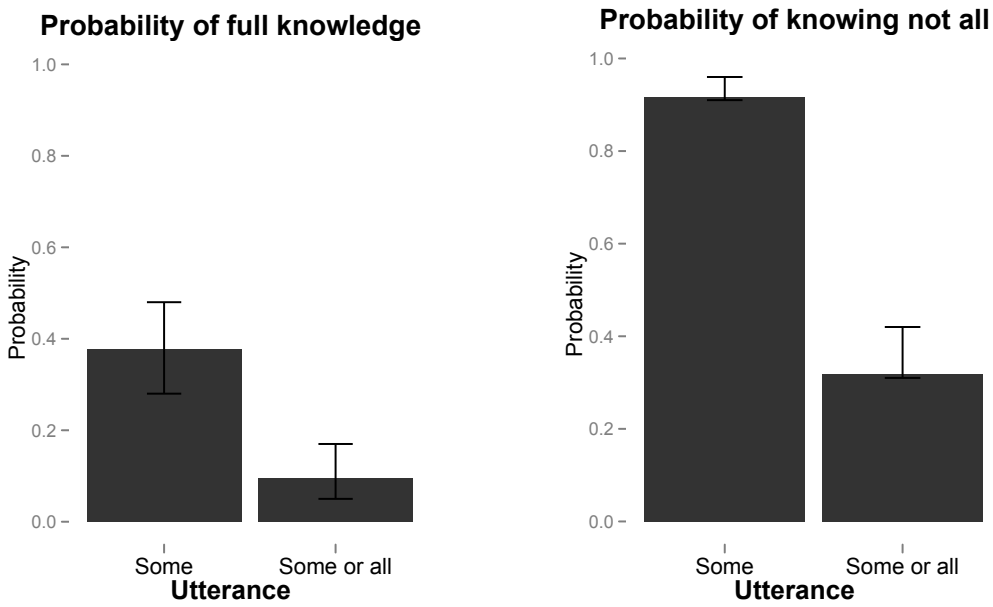
After reading an item, participants were asked two questions:

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 12 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 12(a).

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.[18] 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 12(b).

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.

18 The model which included random slopes for participants did not converge.

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 *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:

$$L_0^{unc}(o, w|u) \propto \sum_{\mathcal{L}'} P(\mathcal{L}')P(o, w)\mathcal{L}'(u, w) \tag{58}$$

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 (58). Define the lexicon $\mathcal{L}_P$ as follows:

$$\mathcal{L}_P(u, w) = \sum_{\mathcal{L}} P(\mathcal{L})\mathcal{L}(u, w) \tag{59}$$

Then it follows from equation (58) that:

$$L_0^{unc}(o,w|u) \propto \sum_{\mathcal{L}'} P(\mathcal{L}')P(o,w)\mathcal{L}'(u,w) \tag{60}$$

$$= P(o,w)\sum_{\mathcal{L}'} P(\mathcal{L}')\mathcal{L}'(u,w) \tag{61}$$

$$= P(o,w)\mathcal{L}_P(u,w) \tag{62}$$

$$\propto L_0(o,w|u,\mathcal{L}_P) \tag{63}$$

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,

$$L_0(o,w|u,\mathcal{L}_P) \propto P(o,w)\mathcal{L}_P(u,w) \tag{64}$$

$$= P(o,w)\sum_{\mathcal{L}'} P(\mathcal{L}')\mathcal{L}'(u,w) \tag{65}$$

$$= P(o,w)\sum_{\mathcal{L}'} P(f(\mathcal{L}'))f(\mathcal{L}')(u',w) \tag{66}$$

$$= P(o,w)\mathcal{L}_P(u',w) \tag{67}$$

$$\propto L_0(o,w|u',\mathcal{L}_P) \tag{68}$$

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:

$$U_1(u|o) = \mathbb{E}_{P_o} \log \frac{1}{L_a(\cdot|u)} - c(u) \tag{69}$$

where L_a is defined by:

$$L_a(\cdot|u) = \textstyle\sum_{\mathcal{L}} P(\mathcal{L}) L_0(\cdot|u, \mathcal{L}) \tag{70}$$

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 (70), $L_a(\cdot|u) = L_a(\cdot|u')$.*

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:

$$L_a(o, w|u) = \sum_{\mathcal{L}} P(\mathcal{L}) L_0(o, w|u, \mathcal{L}) \tag{71}$$

$$= \sum_{\mathcal{L}} P(\mathcal{L}) \frac{P(o, w)\mathcal{L}(u, w)}{Z_{u,\mathcal{L}}} \tag{72}$$

$$= \sum_{\mathcal{L}} P(f(\mathcal{L})) \frac{P(o, w) f(\mathcal{L})(u', w)}{Z_{u', f(\mathcal{L})}} \tag{73}$$

$$= \sum_{\mathcal{L}} P(f(\mathcal{L})) L_0(o, w|u', f(\mathcal{L})) \tag{74}$$

$$= L_a(o, w|u') \tag{75}$$

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.

C Refined lexica for non-convex disjunctions

For reference, Figure 13 depicts the 21 refined lexica in our formalization of Example (52) in Section 5.7.

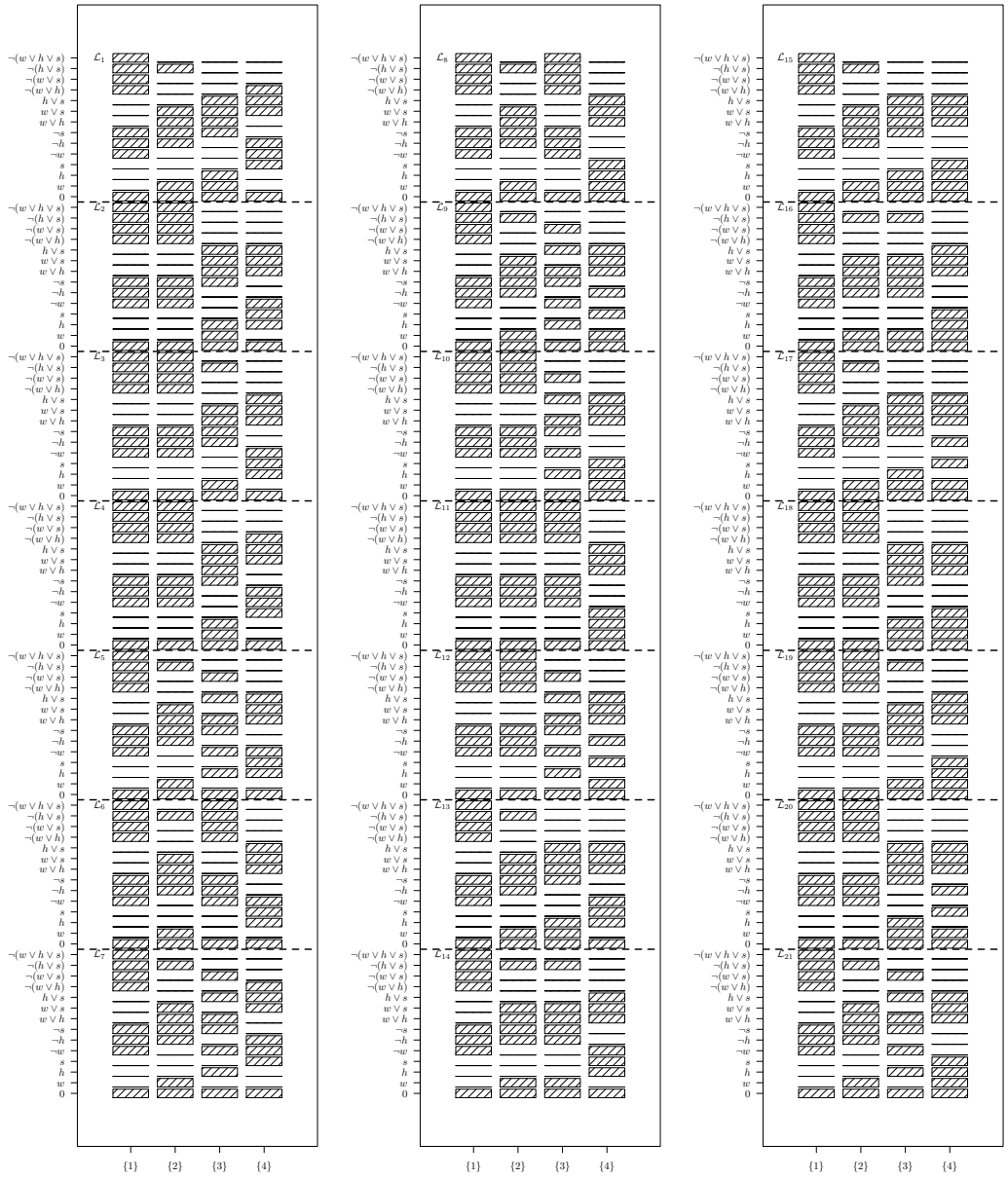


Figure 13 The 21 refined lexica for formalization of Example (52) in Section 5.7.

D Irrelevance of unobserved dimensions

In this section we will show that, if the speaker is presumed to lack knowledge about a dimension of the world, then the lexical uncertainty model predicts that the listener will not gain any information about this dimension from the speaker's utterances. We assume that each world is specified by a vector $(x, y) \in \mathrm{X} \times \mathrm{Y}$.[19] Suppose that the semantic content of each utterance u conveys no information about the dimension Y, that is, if $(x, y) \in [\![u]\!]$, then for all $y' \in Y$, $(x, y') \in [\![u]\!]$. Suppose further that the speaker's observations only provide information about dimension X: for each observation o, $P((x, y)|o) = P(\mathrm{X} = x|o)P(\mathrm{Y} = y)$.

Proposition 1. *Given the assumptions above, $L_i(Y = y|u) = P(Y = y)$ for all utterances u, values y of Y, and $i > 0$.*

Proof. By the definition of L_1 and the assumption of lack of speaker knowledgeability,

$$L_1((x, y), o|u) \propto P(o, (x, y)) \sum_{\mathcal{L}} P(\mathcal{L}) S_1(u|o, \mathcal{L}) \tag{76}$$

$$= P(o)P((x, y)|o) \sum_{\mathcal{L}} P(\mathcal{L}) S_1(u|o, \mathcal{L}) \tag{77}$$

$$= P(o)P(\mathrm{X} = x|o)P(\mathrm{Y} = y) \sum_{\mathcal{L}} P(\mathcal{L}) S_1(u|o, \mathcal{L}) \tag{78}$$

$$= P(\mathrm{Y} = y)F(x, o, u) \tag{79}$$

where $F(x, o, u)$ is defined as a function of the values x, o, u:

$$F(x, o, u) = P(o)P(\mathrm{X} = x|o) \sum_{\mathcal{L}} P(\mathcal{L}) S_1(u|o, \mathcal{L}) \tag{80}$$

It follows that

$$L_1(\mathrm{Y} = y|u) = \frac{P(\mathrm{Y} = y) \sum_{x,o,u} F(x, o, u)}{\sum'_y P(\mathrm{Y} = y') \sum_{x,o,u} F(x, o, u)} \tag{80}$$

$$= \frac{P(\mathrm{Y} = y) \sum_{x,o,u} F(x, o, u)}{\sum_{x,o,u} F(x, o, u) \cdot \sum'_y P(\mathrm{Y} = y')} \tag{81}$$

$$= \frac{P(\mathrm{Y} = y)}{\sum'_y P(\mathrm{Y} = y')} \tag{82}$$

$$= P(\mathrm{Y} = y) \tag{83}$$

19 The arguments in this section generalize to worlds specified by an arbitrary number of dimensions.

The proof for listeners L_i, $i > 1$ is similar. □

This proposition says that the listener gains no information about the dimension Y from the speaker's utterances; given any utterance, their posterior distribution over the value of Y is the same as their prior.

Leon Bergen
Brain and Cognitive Sciences, MIT
43 Vassar St #3037
Cambridge, MA 02139
bergen@mit.edu

Roger Levy
Department of Linguistics, UCSD
9500 Gilman Drive #0108
La Jolla, CA 92093
rlevy@ucsd.edu

Noah Goodman
Department of Psychology, Stanford
Jordan Hall, 450 Serra Mall
Stanford, CA 94305
ngoodman@stanford.edu