Joint Inference in Pragmatic Reasoning

by

Leon Bergen

B.A., Swarthmore College (2009)

Submitted to the Department of Brain and Cognitive Sciences
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2016

Author .... Signature redacted

Department of Brain and Cognitive Sciences
April 20, 2016

Certified by .... Signature redacted

Edward A. F. Gibson
Professor of Cognitive Science
Thesis Supervisor

Accepted by Signature redacted ....

Matthew A. Wilson
Sherman Fairchild Professor of Neuroscience and Picower Scholar
Director of Graduate Education for Brain and Cognitive Sciences

Joint Inference in Pragmatic Reasoning

by

Leon Bergen

Submitted to the Department of Brain and Cognitive Sciences
on April 20, 2016, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

A number of recent proposals have used techniques from game theory and Bayesian cognitive science to formalize Gricean pragmatic reasoning [29, 28, 36, 51]. In the first part of this work, we discuss several phenomena which pose a challenge to these accounts of pragmatics: M-implicatures [45] and embedded implicatures which violate Hurford's constraint [49, 16]. While techniques have been developed for deriving M-implicatures, Hurford-violating embedded implicatures pose a more fundamental challenge to the models' architecture. In order to explain these phenomena, we propose that 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 [36]. The addition of *lexical uncertainty* derives both M-implicatures and the relevant embedded implicatures. This principle explains a novel class of implicature, non-convex disjunctive implicatures. 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 [46, 27]

In the second part of the thesis, we combine these pragmatic models with another recent probabilistic approach to natural language understanding, exploring the formal pragmatics of communication on a noisy channel. We extend a model of rational communication between a speaker and listener, to allow for the possibility that messages are corrupted by noise. Prosodic stress is modeled as the choice to intentionally reduce the noise rate on a word. We show that the model derives several well-known changes in meaning associated with stress, including exhaustive interpretations, scalar implicature strengthening, the association between stress and disagreement, and the interpretation of the focus-sensitive adverbs. We then show that it can account for several phenomena which are outside of the scope of previous accounts of stress interpretation: the effects of stress on quantifier domain inferences, the intensification of gradable adjective interpretation, and the strengthening of hyperbolic utterances. The account avoids the use of syntactic or semantic representations of stress; the interpretive effects of stress are derived from general-purpose pragmatic reasoning.

Thesis Supervisor: Edward A. F. Gibson
Title: Professor of Cognitive Science

Acknowledgments

This work would not have been possible without the support of my mentors. Ted Gibson's focus on important questions, and courage to pursue these questions, have been a model for me. Noah Goodman improved every idea in this work. Josh Tenenbaum provided a vision for the field. All of them have given indispensable advice throughout my Ph.D.

I am grateful to have had the opportunity to collaborate on projects with Roger Levy, Tim O'Donnell, Ev Fedorenko, and Justine Kao; I have learned a great deal from them. Rebecca Saxe provided crucial discussions about this work. Vera Demberg and Guy Van den Broeck were extremely generous in sharing their code and time. I am grateful to many others for their feedback on this work and other projects: Chris Potts, Dan Lassiter, Judith Degen, Greg Scontras, Mike Frank, Michael Franke, Danny Fox, Irene Heim, Nathaniel Smith, Roni Katzir, Benjamin Spector, Tom Wasow, and Kai von Fintel.

The members of TedLab, CoCoSci, and CoCoLab, and many others in BCS, have made my Ph.D. much better than it would have been otherwise: Steve Piantadosi, Richard Futrell, Kyle Mahowald, Melissa Kline, Andreas Stuhlmuhller, Owain Evans, Max Siegel, Owen Lewis, Jon Malmaud, Sam Norman-Haignere, Michael Gelbart, Tomer Ullman, John McCoy, Sam Gershman, Eyal Dechter, John Kraemer, Chris Baker, Julian Jara-Ettinger, Zenna Tavares, Long Ouyang, Michael Henry Tessler, and Erin Bennett.

I am lucky to have had many great teachers before getting to MIT. I would in particular like to thank Tim Corica and Dan Grodner; this work might not have existed if I had not met them at the right times. This work certainly would not have existed with my parents and my Aunt Libby and Uncle Lowell, who have been dedicated to my education for nearly three decades.

Finally, I would like to thank Karen, for her brilliance, her kindness, and her love. I owe everything to her.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

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

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

Despite the promise and apparently broad empirical coverage of these theories, attempts at formalizing them (e.g., 33) have historically met with less success than formalization in other linguistic domains such as phonology, syntax, or semantics. Nevertheless, there is strong reason to believe that formal accounts of Gricean pragmatic reasoning have substantial potential scientific value. First, all Gricean theories assume that multiple factors—most famously Grice's quality, quantity, relevance, and manner—jointly guide the flexible relationship between literal semantic content and understood meaning, and in all Gricean theories these factors can potentially come into conflict (e.g., the opposition between Horn's [45] 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., 45, 62, 92]. 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 [69, 91] and cognitive science [96]. 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 [77, 29, 51, 31, 85, 28, 36, 22, 6, 87]. 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 [28, 36, 22].

This work 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.

The last phenomenon that we consider in the thesis's first section is 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 — i.e., atomic expressions — rather than whole expressions. When the model is extended in this manner, it will derive the embedded implicatures in question.

The phenomena discussed in the first section 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 in , 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.

In the second part of the thesis, we explore the formal pragmatics of communication over a noisy channel. Recent work in cognitive science has provided evidence that people rationally account for the possibility of noise in their language input [2, 64, 34]. We extend the rational speech acts model, so that the speaker and listener rationally adjust for the possibility of noise, and so that this fact is *common knowledge* between them. This framework allow us to model prosodic stress as the intentional reduction of the noise rate on part of the utterance. The speaker in this model exploits the possibility of noise, and uses stress to communicate aspects of their intended meaning. We show that the model derives several well-known changes in meaning associated with stress, including exhaustive interpretations, scalar implicature strengthening, the association between stress and disagreement, and the interpretation of the focus-sensitive adverbs. We then show that it can account for several phenomena which are, to the best of our knowledge, outside of the scope of previous accounts of stress interpretation: the effects of stress on quantifier domain inferences, the intensification of gradable adjective interpretation, and the strengthening of hyperbolic utterances. The account avoids the use of syntactic or semantic representations of stress; all of the interpretive effects of stress are derived in a strictly pragmatic manner.

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 work 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 thesis: 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.

Chapter 2

Pragmatic Reasoning through Semantic Inference

2.1 The baseline rational speech-act theory of pragmatics

We begin by introducing the baseline rational speech-act theory of pragmatics [28, 36], built on a number of simple foundational assumptions about speakers and listeners in cooperative communicative contexts. We assume first a notion of COMMON KNOWLEDGE [65, 94, 18]—information known by both speaker and listener, with this shared knowledge jointly known by both speaker and listener, knowledge of the knowledge of shared knowledge jointly known by both speaker and listener, and so on *ad infinitum* (or at least as many levels of recursion up as necessary in the recursive pragmatic inference). Communication involves the transmission of knowledge which is not common knowledge: we assume that the speaker, by virtue of some observation that she has made, is in a particular belief state regarding the likely state of some conversationally relevant aspect of the world (or, more tersely, regarding the world). In engaging in a cooperative communicative act, the speaker and listener have the joint goal of bringing the listener's belief state as close as possible to that of the speaker, by means of the speaker formulating and sending a not-too-costly signal to the listener, who interprets it. The lexicon and grammar of the speaker and listener's language serve as resources by which literal content can be formulated. As pragmatically sophisticated agents, the speaker and the listener recursively model each other's expected production decisions and inferences in comprehension.

More formally, let O be the set of possible speaker observations, $\mathcal{W}$ the set of possible worlds, and $\mathcal{U}$ the set of possible utterances. Observations $o \in O$ and worlds $w \in \mathcal{W}$ have joint prior distribution $P(o,w)$, shared by listener and speaker.

The literal meaning of each utterance $u \in \mathcal{U}$ is defined by a lexicon $\mathcal{L}$, which is a mapping from each possible utterance-world pair to the truth value of the utterance in that world. That is,

$$\mathcal{L}(u,w) = \begin{cases} 0 & \text{if } w \notin [\![u]\!] \\ 1 & \text{if } w \in [\![u]\!] \end{cases} \tag{2.1}$$

where $[\![u]\!]$ is the intension of u.[1]

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

$$L_0(o,w|u,\mathcal{L}) \propto \mathcal{L}(u,w)P(o,w). \tag{2.2}$$

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,

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

corresponding to whether the speaker observed that all of the students passed ($\forall_o$), observed

[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 2.4.

that some but not all of them passed ($\exists\neg\forall_o$), or did not make any relevant observations ($\emptyset_o$). 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

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

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:

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

Only two observation-world pairs are 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 2.2. 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

$$U_n(u|o) = \mathbb{E}_{P(w|o)} \log L_{n-1}(o,w|u) - c(u) \tag{2.3}$$

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 work, 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:

$$\mathbb{E}_{P(w|o)} \log L_{n-1}(o,w|u) = \sum_{w} P(w|o) \log L_{n-1}(o,w|u) \tag{2.4}$$

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 information 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 2.4 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 work, we assume that for each world $w \in \mathcal{W}$, there is a unique observation $o \in O$ consistent with this world. In this special case, it is common knowledge that the speaker knows the true world w with probability 1, so that $P(w|o)$ is 1 for that world and 0 for all other worlds. This entails that we can ignore the world variable w in the speaker and listener equations, and the 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

$$U_n(u|o) = \log L_{n-1}(o|u) - c(u) \tag{2.5}$$

The assumption of speaker knowledgeability is relaxed in Section 2.3.6.

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

$$S_n(u|o) \propto e^{\lambda U_n(u|o)}, \tag{2.6}$$

where $\lambda > 0$. This specification of the speaker formula uses the SOFTMAX FUNCTION or LUCE-CHOICE RULE [95] 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 [95].

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

$$L_n(o,w|u) \propto P(o,w)S_n(u|o). \tag{2.7}$$

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 (2.2), (2.3), (2.6), and (2.7) constitute the heart of this basic model. Note the relationship between recursion levels of the speaker and listener in Equations (2.3): 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 [36] by adding utterance costs and the possibility of recursion beyond S_1.

2.1.1 Auxiliary assumptions: alternative sets, but no lexical scales

As in much previous work in pragmatics [37, 45, 33, 62], our models of pragmatic reasoning will rely heavily the set of alternative utterances available to the speaker. That is, in deriving the implicatures for an utterance, our models will reason about why the speaker did not use the other utterances available to them. We will not be providing a general theory of the alternative utterances that are reasoned about during the course of pragmatic inference. Rather, as is done in most other work in pragmatics, we will posit the relevant set of utterances on a case-by-case basis. As is discussed below, however, there are certain cases for which our models require fewer restrictions on the set of alternatives than most other models. These examples will provide suggestive — though not decisive — evidence that no categorical restrictions need to be placed on the alternatives set within our models, i.e. that every grammatical sentence in a language can be considered as an alternative during pragmatic reasoning. The mechanisms by which this may be made possible are discussed below.

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

2.2 Specificity implicature in the baseline theory

To demonstrate the value of the baseline theory presented in Section 2.1, 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 28 for experimental evidence for these implicatures). That is, in this context, there is an available utterance which is more specific than “green", and as a result “green" receives a specificity implicature which is the negation of the more specific utterance. It is important to note that neither "green" nor "ball" is strictly logically stronger than the other; it is only in a particular context that one can be strictly more descriptive than the other. Thus, these utterances do not fall on a scale which is ordered according to logical strength.[2]

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

2.2.1 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,

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

[2] Though these utterances are logically incommensurable, it may still be possible to describe them as falling on an *ad-hoc* scale, as in [42]. 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.

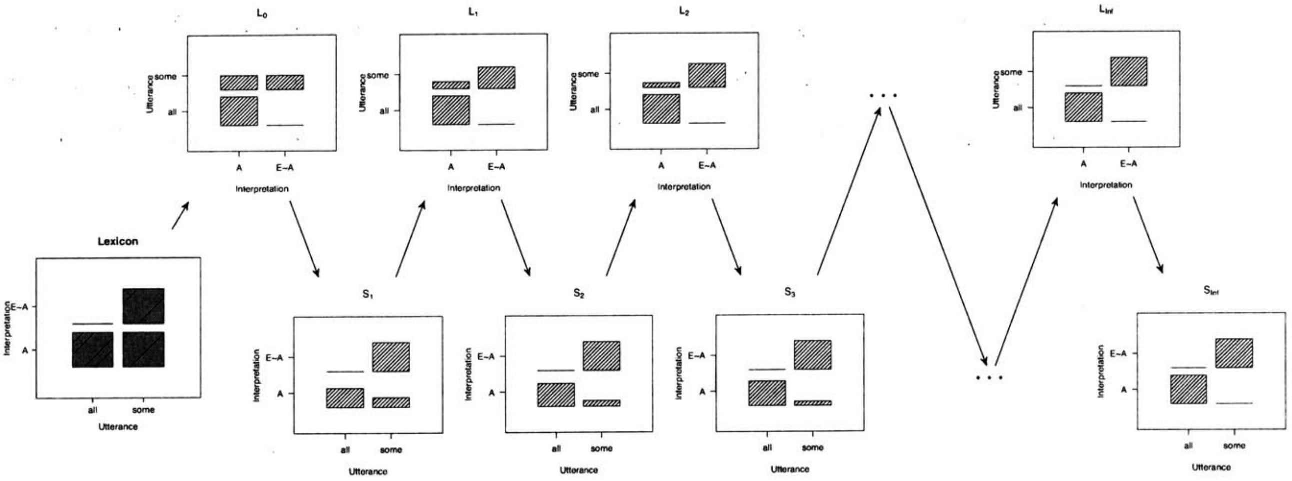


Figure 2-1: *Some* strengthening with $\mathrm{P}(\forall) = \frac{1}{2}$, $\mathrm{P}(\exists\neg\forall) = \frac{1}{2}$, $c(\text{all}) = c(\text{some}) = 0$, $\lambda = 1$. The lexicon panel indicates the truth value of utterances across each world. The listener panels indicate the conditional probabilities 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 this work, are intended for readers who want to better understand the dynamics of the speaker-hearer recursion. The linguistic claims of this work can be appreciated without relying on them.

and two meanings,

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

where the intensions of the utterances are as usual:

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

Since it is common knowledge that the speaker knows the relevant world state, we can without loss of generality consider the observation and world variables to be equal, so that $o = w$, and drop w from the recursive equations (2.2)– (2.7). This allows the baseline model to be expressed as

$$L_0(o|u,\mathcal{L}) \propto \mathcal{L}(u,o)P(o), \tag{2.8}$$

$$U_n(u|o) = \log L_{n-1}(o|u) - c(u), \tag{2.9}$$

$$S_n(u|o) \propto e^{\lambda U_n(u|o)}, \tag{2.10}$$

$$L_n(o|u) \propto P(o)S_n(u|o), \tag{2.11}$$

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 $\lambda = 1$.[3]

Figure 2-1 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):

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

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

[3] Changes in the prior on observations, utterance costs, and the softmax parameter change the precise values of the speaker and listener posteriors at various levels of recursion, but do not change the signature specificity-implicature pattern that the model exhibits. For this example, and for others throughout this work, we assessed robustness to changes in the model parameters by computing model predictions across a grid of parameter values.

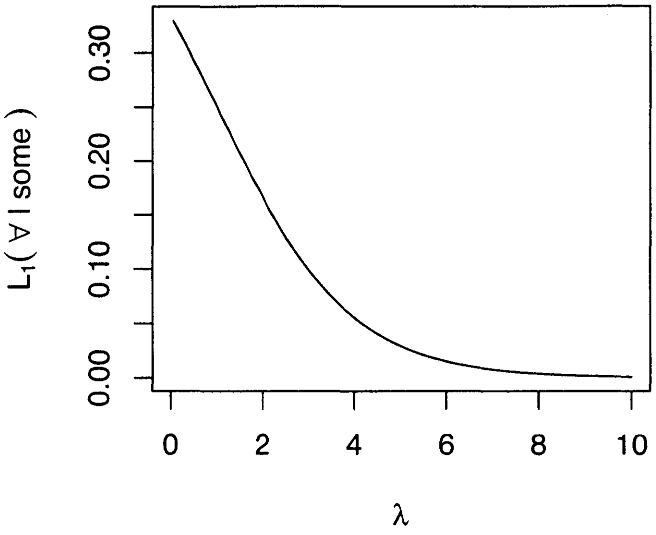


Figure 2-2: The degree of *some* strengthening as a function of the "greedy rationality" parameter λ, with $P(\forall) = \frac{1}{2}$, $P(\exists\neg\forall) = \frac{1}{2}$, $c(\text{all}) = c(\text{some}) = 0$

One way to understand the fundamental reason for this behavior—the signature pattern of specificity implicature—is by considering the effect on one level of recursive inference on the listener's tendency to interpret *some* with unstrengthened meaning $\forall$. Let us denote $L_{n-1}(\forall|\text{some})$ by the probability p. Further, note that lexical constraints on the literal listener mean that $L_n(\exists\neg\forall|\text{all}) = 0$ always. This means that we can write, following Equations (2.9)–(2.11):

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 work, 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 2-2 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.

2.2.2 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, i.e. that "all" is true.

[26] 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.

2.3 Lexical uncertainty

2.3.1 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:

(1) John can finish the homework.

(2) 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. [45] and [62] 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. [7] find that in one-shot communication games of this sort, people do in fact communicate efficiently, suggesting that the pragmatic knowledge underlying M-implicatures is quite general and not limited to specific linguistic examples.[4]

Failure of rational speech acts model to derive M-implicatures

Perhaps surprisingly, our baseline rational speech-act model of Sections 2.1–2.2 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 2-3; the model's qualitative failure is totally general across different settings of prior probabilities, utterance costs, and λ. The literal listener L_0 interprets both utterances identically, following the prior probabilities of the meanings. Crucially, L_0's interpretation distribution provides no information that speaker S_1 can leverage to associate either utterance with any specific meaning; the only thing distinguishing the utterances' expected utility is their cost. This leads to an across-the-board dispreference on the part of S_1 for *long*, but gives no starting point for more sophisticated listeners or speakers to break the symmetry between these utterances.

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

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

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

Proof. By equation 2.2,

$$L_0(o,w|u,\mathcal{L}) = \frac{P(o,w)\mathcal{L}(u,w)}{\sum_{o',w'} P(o',w')\mathcal{L}(u,w')} \tag{2.12}$$

$$= \frac{P(o,w)\mathcal{L}(u',w)}{\sum_{o',w'} P(o',w')\mathcal{L}(u',w')} \tag{2.13}$$

$$= L_0(o,w|u',\mathcal{L}) \tag{2.14}$$

where the equality in 2.13 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 2.3,

$$U_n(u|o) - c(u') = \mathbb{E}_{P(w|o)} \log L_{n-1}(o,w|u) - c(u) - c(u') \tag{2.15}$$

$$= \mathbb{E}_{P(w|o)} \log L_{n-1}(o,w|u') - c(u') - c(u) \tag{2.16}$$

$$= U_n(u'|o) - c(u) \tag{2.17}$$

It follows from equation 2.6 that:

$$S_n(u|o) = \frac{e^{\lambda U_n(u|o)}}{\sum_{u_i} e^{\lambda U_n(u_i|o)}} \tag{2.18}$$

$$= \frac{e^{\lambda(U_n(u'|o)-c(u)+c(u'))}}{\sum_{u_i} e^{\lambda U_n(u_i|o)}} \tag{2.19}$$

$$= S_n(u'|o) \cdot e^{\lambda(c(u')-c(u))} \tag{2.20}$$

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

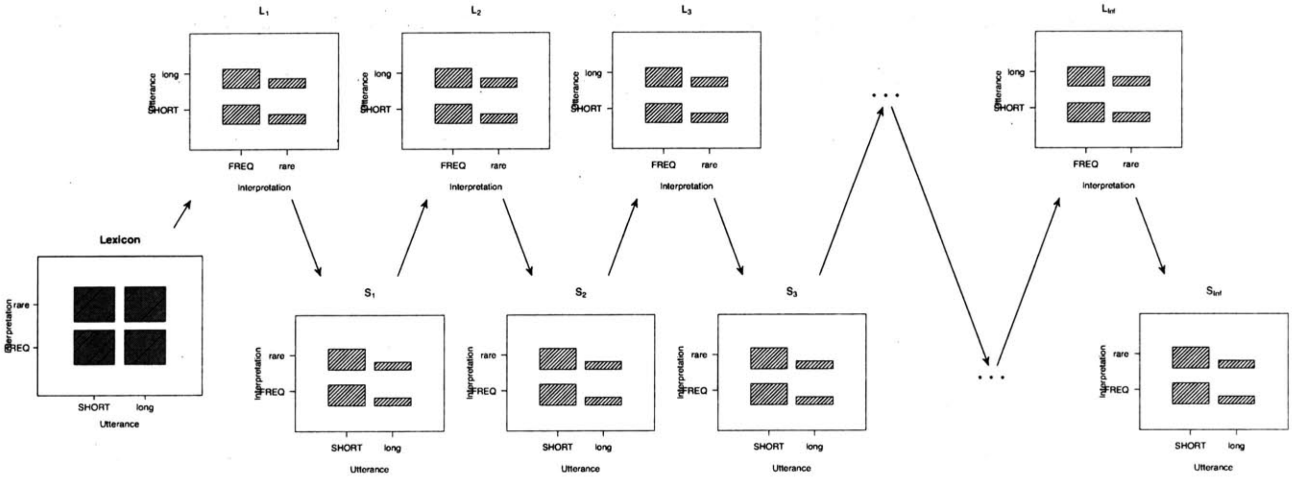


Figure 2-3: The failure of the basic model to derive M-implicature (illustrated here for $P(\text{FREQ}) = \frac{2}{3}$, $P(\text{rare}) = \frac{1}{3}$, $\lambda = 3$, $c(\text{SHORT}) = 1$, $c(\text{long}) = 2$). The listener panels illustrate that the interpretation of each utterance is constant across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

We will now show the equivalence of listeners $L_n(\cdot|u)$ and $L_n(\cdot|u')$. By equation 2.7,

$$L_n(o,w|u) = \frac{P(o,w)S_n(u|o)}{\sum_{o',w'} P(o',w')S_n(u|o')} \tag{2.21}$$

$$= \frac{P(o,w)S_n(u'|o)e^{\lambda(c(u')-c(u))}}{\sum_{o',w'} P(o',w')S_n(u'|o')e^{\lambda(c(u')-c(u))}} \tag{2.22}$$

$$= L_n(o,w|u') \tag{2.23}$$

□

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.

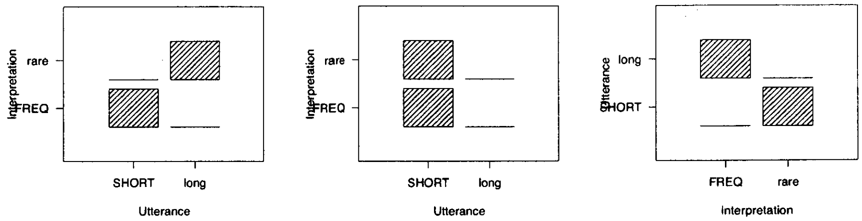
2.3.2 The multiple equilibrium problem

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 [79, 17]. 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 work, each speaker has a type, which is the meaning that they want to communicate. The goal of the listener is to correctly guess the type of the speaker based on the signal that they send.

To describe the multiple equilibrium problem for such games, we first need to introduce the relevant notion of equilibrium. Loosely speaking, the equilibria for a game describe the self-consistent ways that the game can be played. The simplest equilibrium concept in game theory is the Nash equilibrium [76, 74, 32]. 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 [41], which in addition to the requirements imposed by the definition of the Nash equilibrium also imposes consistency constraints on the beliefs of the agents. In particular, given the prior distribution over types, and the agents' strategies (which define the likelihood of taking actions given a player type), the agents must use Bayes' rule to compute their posterior distribution over types after observing an action. Each agent's strategy must also be rational given their beliefs at the time that they take the action, in the sense that the strategy must maximize their expected utility. The multiple equilibrium problem arises in a signalling game when the game has multiple Bayesian Nash equilibria. This occurs when the agents can devise multiple self-consistent communication systems given the constraints of the game. That is, given the assumption that the other agents are using the communication system, it will not be rational for one agent to unilaterally start using a different communication system.

The multiple equilibrium problem can be illustrated concretely using the game above. This game has two general classes of equilibria, illustrated in Figure 2-4. 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



(a) Separating Pareto-optimal

(b) Pooling 1

(c) Separating Pareto-suboptimal

Figure 2-4: Multiple equilibria (speaker matrices) for the M-implicature signaling game

long when they want to communicate *rare*, and SHORT when they want to communicate FREQ (Figure 2-4a). Given these strategies, the listener knows how to interpret each utterance: *long* will be interpreted as *rare*— conditional on hearing *long*, the only possibility is that it was produced by the agent wanting to communicate *rare*— and similarly SHORT will be interpreted as FREQ. This is clearly an equilibrium, because neither speaker will successfully communicate their intended meaning if they unilaterally change their strategy; for example, if the speaker wanting to communicate *rare* switches to using SHORT, then they will be interpreted as intending FREQ. A second separating equilibrium is also possible in this game. Under this equilibrium, the speaker-utterance pairs are reversed, so that the agent intending to communicate *rare* uses SHORT, and the agent intending FREQ uses *long* (Figure 2-4c). This is inefficient — in expectation, it will be more expensive than the previous equilibrium for the speaker — but it is nonetheless an equilibrium, because neither speaker can unilaterally change strategies without failing to communicate.

The second type of equilibrium in this game, known as a *pooling equilibrium*, is still more deficient than the inefficient separating equilibrium, and it is the one that is most closely related to the problem for our initial model of pragmatic inference. In one pooling equilibrium, the speaker chooses the utterance SHORT, independent of the meaning that they want to communicate (Figure 2-4b). Because the speakers always choose SHORT, this utterance communicates no information about the speaker's intended meaning, and the listener interprets this utterance according to the prior distribution on meanings. Assuming that the utterance *long* is also interpreted according to the prior, it will never be rational for the speaker to choose this utterance.[5] Thus this is indeed an equilibrium.

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

In recent work in linguistics, there have generally been three approaches to accounting for these reasoning abilities. The first approach uses the notion of a *focal point* for equilibria [77]. On this approach, people select the efficient equilibrium in signalling games because it is especially salient; the fact that it is salient makes each agent expect other agents to play it, which in turn makes each agent more likely to play it themselves. While this approach does derive the efficient equilibrium for communication games, it is not entirely satisfactory, since it does not provide an independent account of salience in these games — precisely the feature which allows the agents to efficiently communicate under this approach.

An alternative approach has been to derive the efficient equilibrium using evolutionary game theory, as in [97, 21]. 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

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

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, [29], [51], and [31] 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. [29] and [51] 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. [31] therefore provide an alternative extension of the IQR model which derives M-implicatures. Under this proposal, agents who receive low utility from all of their available actions engage in more exploratory behavior. In a Horn game, the speaker who wants to communicate the meaning *rare* starts out with a low expected utility from all of their actions: no matter which utterance they choose, the listener is unlikely to interpret them correctly. As a result, this speaker will engage in more exploratory behavior — i.e., behave less optimally with respect to their communicative goal — and will be more likely to choose the suboptimal expensive utterance. This is sufficient to break the symmetry between the cheap and expensive utterances, and derive the M-implicature.

Unlike the proposed modification of the IBR model, [31]'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 ıational 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.

2.3.3 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 A.2 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 2.2, i.e. this listener is defined by:

$$L_0(o,w|u,\mathcal{L}) \propto \mathcal{L}(u,w)P(o,w) \tag{2.24}$$

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:

$$\begin{aligned}
U_1(u|o,\mathcal{L}) &= \mathbb{E}_{P(w|o)} \log L_0(o,w|u,\mathcal{L}) - c(u), && && (2.25)\\
S_1(u|o,\mathcal{L}) &\propto e^{\lambda U_1(u|o,\mathcal{L})}, && && (2.26)\\
L_1(o,w|u) &\propto P(o,w) \sum_{\mathcal{L}\in\Lambda} P(\mathcal{L})S_1(u|o,\mathcal{L}), && && (2.27)\\
U_n(u|o) &= \mathbb{E}_{P(w|o)} \log L_{n-1}(o,w|u) - c(u) && \text{for } n > 1, && (2.28)\\
S_n(u|o) &\propto e^{\lambda U_n(u|o)} && \text{for } n > 1, && (2.29)\\
L_n(o,w|u) &\propto P(o,w)S_n(u|o) && \text{for } n > 1.^6 && (2.30)
\end{aligned}$$

These equations differ from the baseline model in several respects. In Equations 2.25 and 2.26, 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 2.27 is the most important difference between the current and baseline models. The listener L_1 in the baseline model (Equation 2.7) is certain about the speaker S_1's beliefs about the lexicon; for this listener, there is a single lexicon which determines the speaker's beliefs about the literal meanings of utterances. In the current model, the

[6] It is possible to define the lexical-uncertainty model more concisely by replacing Equations (2.25)–(2.30) with the following three equations:

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

Once the first marginalization over lexica occurs at the L_1 level, higher-level speaker and listener distributions lose their dependence on the lexicon $\mathcal{L}$ being conditioned on, since there is no dependence on $\mathcal{L}$ in the right-hand side of equation (iii). In this work 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).

listener L_1 (Equation 2.27) 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 2.28-2.30, are the same as in the baseline model.

In order for the normalization of Equation 2.24 and the expected surprisal of Equation 2.25 to be well-defined we must place two restrictions on each $\mathcal{L} \in \Lambda$.

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

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

Satisfying the first of these restrictions is straightforward. We have considered four approaches to constructing Λ that satisfy the second restriction, each of which result in qualitatively similar predictions for all of the models considered in this work. In the first of these approaches, the global constraint of restriction 2 is simply imposed on each lexicon by fiat; any lexicon which does not satisfy this condition is assigned probability 0. In the second of these approaches, the truth-conditional semantics of each utterance is slightly weakened. When an utterance u is false at a world state w, we define $\mathcal{L}(u,w) = 10^{-6}$ (or any smaller, positive number). In this case, each utterance always assigns at least a small amount of mass to each world state, immediately satisfying restriction 2. In the third approach we assume that there is some, much more complex, utterance that could fully specify any possible belief state. That is, for any o there is an utterance u_o such that $\mathcal{L}(u_o,\cdot)$ coincides with the support of $P(w|o)$ in every lexicon $\mathcal{L} \in \Lambda$. The utterances u_o may be arbitrarily expensive, so that the speaker is arbitrarily unlikely to use them; they still serve to make the 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 work, 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 (i.e. the lexicon that maps each utterance to its truth function under the language's semantics) we can define Λ by a canonical procedure of sentential enrichment: Call the utterance meaning $\mathcal{L}(u, \cdot)$ a *valid refinement* of $\mathcal{L}_S$ if: $\forall w\ \mathcal{L}_S(u,w) = 0 \implies \mathcal{L}(u,w) = 0$, and, $\exists w\ \mathcal{L}(u,w) > 0$. 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.

2.3.4 Specificity implicature under lexical uncertainty

Before demonstrating how the lexical-uncertainty model derives M-implicature (which we do in Section 2.3.5), 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 $\mathcal{U} = \{\text{some, all}\}$, meanings $\mathcal{W} = \{\forall, \exists\neg\forall\}$, and literal utterance meanings—semantic lexicon $\mathcal{L}_S$ in the terminology of Section 2.3.3—

$[\![\text{some}]\!] = \{\forall, \exists\neg\forall\}$, $[\![\text{all}]\!] = \{\forall\}$. The enrichment procedure gives Λ consisting of:

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

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

We can now compute the behavior of the model. Since it is common knowledge that the speaker knows the relevant world state, we can once again let $o = w$ and drop w from the recursive equations, so that the lexical-uncertainty model of Equations (2.24)–(2.30) can be expressed as

$$L_0(o|u,\mathcal{L}) \propto \mathcal{L}(u,o)P(o), \tag{2.31}$$

$$U_1(u|o,\mathcal{L}) = \log L_0(o|u,\mathcal{L}) - c(u), \tag{2.32}$$

$$S_1(u|o,\mathcal{L}) \propto e^{\lambda U_1(u|o,\mathcal{L})}, \tag{2.33}$$

$$L_1(o|u) \propto P(o) \sum_{\mathcal{L}\in\Lambda} P(\mathcal{L})S_1(u|o,\mathcal{L}), \tag{2.34}$$

$$U_n(u|o) = \log L_{n-1}(o|u) - c(u) \qquad \text{for } n > 1, \tag{2.35}$$

$$S_n(u|o) \propto e^{\lambda U_n(u|o)} \qquad \text{for } n > 1, \tag{2.36}$$

$$L_n(o|u) \propto P(o)S_n(u|o) \qquad \text{for } n > 1. \tag{2.37}$$

Figure 2-5 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.

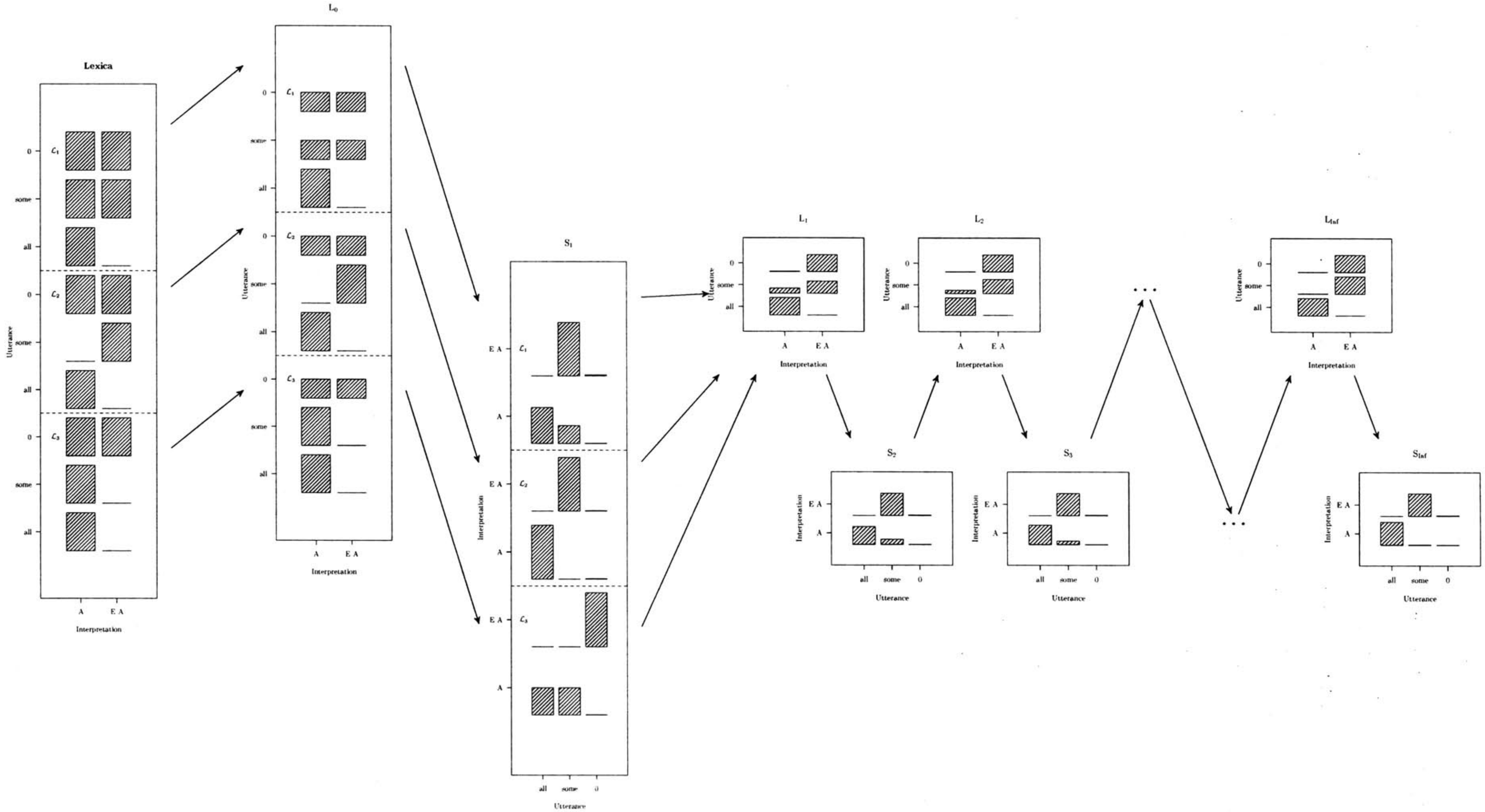


Figure 2-5: Specificity implicatures under lexical uncertainty, shown here with $P(\forall) = \frac{1}{2}$, $P(\exists\neg\forall) = \frac{1}{2}$, $c(\text{all}) = c(\text{some}) = 1$, $c(\mathbf{0}) = 5$, $\lambda = 1$. The first column shows the three admissible lexica in this example. The second column 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.

The specificity implicature first appears at the level of the listener L_1, who is reasoning about the speaker S_1. The listener computes their posterior distribution over the speaker's intended meaning by marginalizing over the possible lexica that the speaker may have been using (Equation (2.34)). As can be seen in the second column of the third panel of Figure 2-5, 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.

2.3.5 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—$\{\text{FREQ}, \textit{rare}\}$, $\{\text{FREQ}\}$, and $\{\textit{rare}\}$—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 2-6, 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 2-6). 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 (2.34)), 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 2-6 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 [7]).

2.3.6 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:

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

(4) Some of the students passed the test.

As discussed in Section 2.2, (4) has a specificity implicature that strengthens the literal meaning of "some" to an understood meaning of "some but not all". The implicatures of (3) differ crucially in two ways. First, as noted by [33, see also 16], (3) lacks the basic specificity implicature of (4). Second, (3) 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

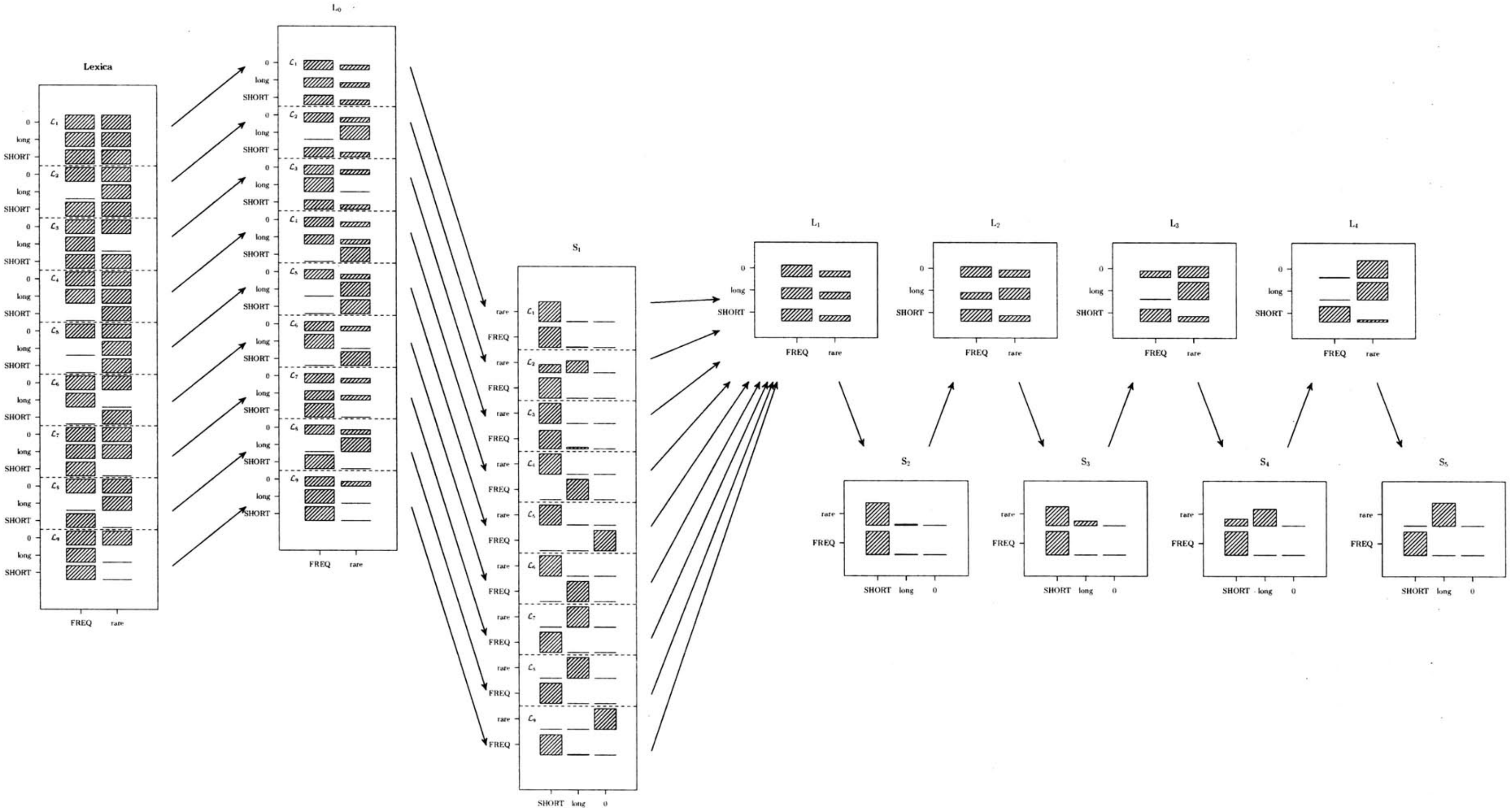


Figure 2-6: Deriving M-implicatures with $P(\text{FREQ}) = \frac{2}{3}$, $P(\text{rare}) = \frac{1}{3}$, $\lambda = 4$, $c(\text{SHORT}) = 1$, $c(\text{long}) = 2$, $c(\emptyset) = 5$. As in figure 2-5, the first column enumerates all of the admissible lexica, and the next two columns show the listener 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.

and pragmatic theory [24, 87, 71, 25]. This is for several reasons. First, the sentence in (3) violates Hurford's constraint [49], 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 2.1 — cannot derive the implicatures associated with this sentence. A disjunction which violates Hurford's constraint will be semantically equivalent to one of its disjuncts (i.e. the weaker one); in this case, the expression "some or all" is semantically equivalent to "some." As previously discussed, the rational speech acts model, and neo-Gricean models more generally, cannot derive distinct pragmatic interpretations for semantically equivalent expressions. 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 [7].

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 (3) 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 (3) possess an ignorance implicature that corresponding sentences such as (4) do not should make the following empirically testable prediction. Consider a sentence pair like ((3)–(4)), 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

[7] http://www.sfs.uni-tuebingen.de/ gjaeger/slides/slidesIrvine.pdf

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.1.)

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 2.2 and 2.3.4, 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 2.3.3, 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 2-7 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

[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 $P(w|o)$.

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 2.3.5, but rather is a distinct phenomenon. 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.

2.4 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

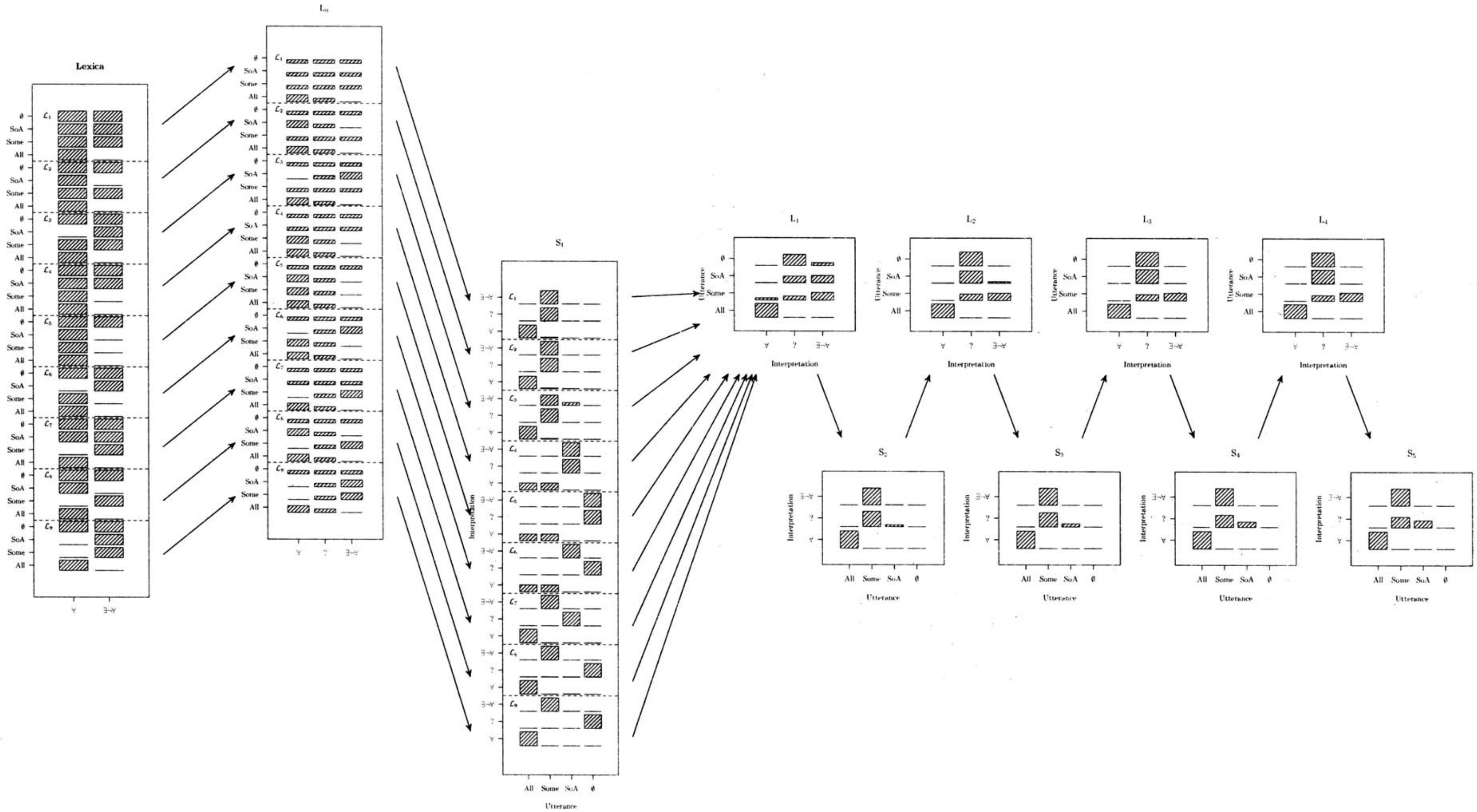


Figure 2-7: Deriving generalized markedness implicatures with λ = 4, uniform prior probabilities over observations, c("all")=0, c("some")=0, c("some or all")=1.

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 work, one could adopt the generalized notion of entailment from natural logics and then define a refinement of a lexical entry as another term of the same type that entails the original entry. The set of lexica Λ could then be derived, as before, as the set of all lexicons that can be derived from $\mathcal{L}_S$ by refinement. Sentence meanings would then be derived from (refined) lexical meanings by ordinary compositional mechanisms. In this work, 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 2.3.3. Our previous solution for restriction 2 carries over: we may extend each lexicon with a trivial u_{null}. Restriction 1 is more subtle than before. We must still guarantee that the literal listener can interpret any utterance. Simply restricting that the lexical entries be assigned non-contradictory refinements in not enough, as composition can arrive an contradictions (e.g. "A and not A"). There are various options available to solve this problem[9]; we will initially restrict our attention to composition by disjunction, where it is sufficient to require that individual

[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 [35].

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.

2.4.1 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* [62], where the relevant notion of markedness is the prior probability of the meaning: ordinary smiles are more common than smirks and grimaces. The second example was that of ignorance implicatures for disjunctions such as *some or all*, in which the relevant notion of markedness is the degree of speaker ignorance about the world state: the more complex utterance is interpreted as indicating a greater degree of speaker ignorance. However, there are even more challenging cases than these: cases in which non-atomic utterances with identical literal content *and* identical formal complexity receive systematically different interpretations. A general class of these cases can be constructed from entailment scales containing more than two items, by creating a disjunction out of two non-adjacent terms on the scale:

(5) 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.*]

(6) 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.*]

(7) 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 alternatives disjunctive expressions with identical formal complexity (in the sense of having the same syntactic structure and number of words) and literal meaning under standard assumptions that the literal meanings of such expressions are lower bounds in the semantic space of the scale, but different understood meaning: *warm or hot, decent or good, one or two.*[10] It is not at all clear on a standard neo-Gricean account how these pairs of alternatives come to have different pragmatic interpretations. Second, these expressions have the property that their understood meanings are NON-CONVEX within the semantic space of the scale. This property poses a serious challenge for standard neo-Gricean accounts: since all the alternatives whose negation could be inferred through pragmatic reasoning have literal meanings that are upper bounds in the semantic space, it is unclear how the resulting pragmatically strengthened meaning of the utterance could ever be non-convex.

The basic lexical uncertainty framework developed in Section 2.3 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

[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. 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.

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.

2.4.2 Compositional lexical uncertainty

In this section we further specify compositional lexical uncertainty, as sketched out above, for the case of boolean atomic utterances composed by disjunction. This requires only a small change to the original lexical-uncertainty model introduced in Section 2.3: 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:

$$[\![u]\!]_{\mathcal{L}}(w) = \mathcal{L}(u,w) \tag{2.38}$$

[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.

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":

$$[\![u_1 \text{or } u_2]\!]_{\mathcal{L}}(w) = \begin{cases} 1 & \text{if } [\![u_1]\!]_{\mathcal{L}}(w) = 1 \text{ or } [\![u_2]\!]_{\mathcal{L}}(w) = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{2.39}$$

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:

$$L_0(w, o | u, \mathcal{L}) \propto [\![u]\!]_{\mathcal{L}}(w) P(w, o) \tag{2.40}$$

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.

2.4.3 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:

$$U_A = \{\textit{one}, \textit{two}, \textit{three}\} \qquad O = \begin{tikzcd} 1 \arrow[d, dash] \arrow[dr, dash] & 2 \arrow[dl, dash] \arrow[dr, dash] & 3 \arrow[dl, dash] \arrow[d, dash] \\ 1 \vee 2 \arrow[dr, dash] & 1 \vee 3 \arrow[d, dash] & 2 \vee 3 \arrow[dl, dash] \\ & 1 \vee 2 \vee 3 & \end{tikzcd}$$

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

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

for the simple indicative case.

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

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

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), 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 2-8 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 $\{2,3\}$). 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

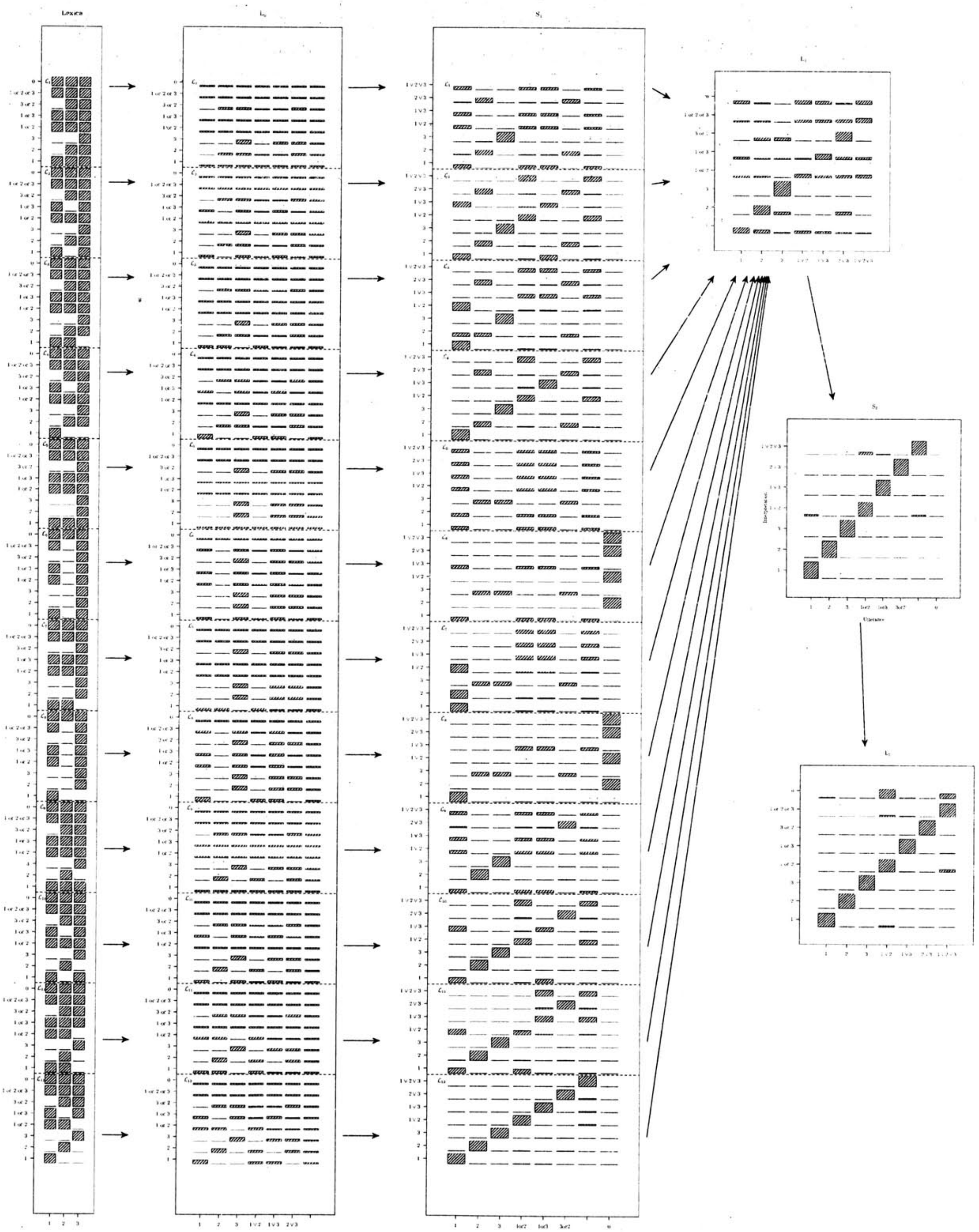


Figure 2-8: Non-convex disjunction, for uniform marginal distribution $P(O)$, uniform conditional distributions $P(W|O)$, cost per disjunct of 0.05, and $\lambda = 5$. Only lexica (and L_0 and S_1 distributions) in which the refined meaning of *one* contains the world state 1 are shown.

inference.

2.4.4 *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 2.3.6, 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]

2.4.5 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 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 2.3.4, 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

[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 2.3.6.

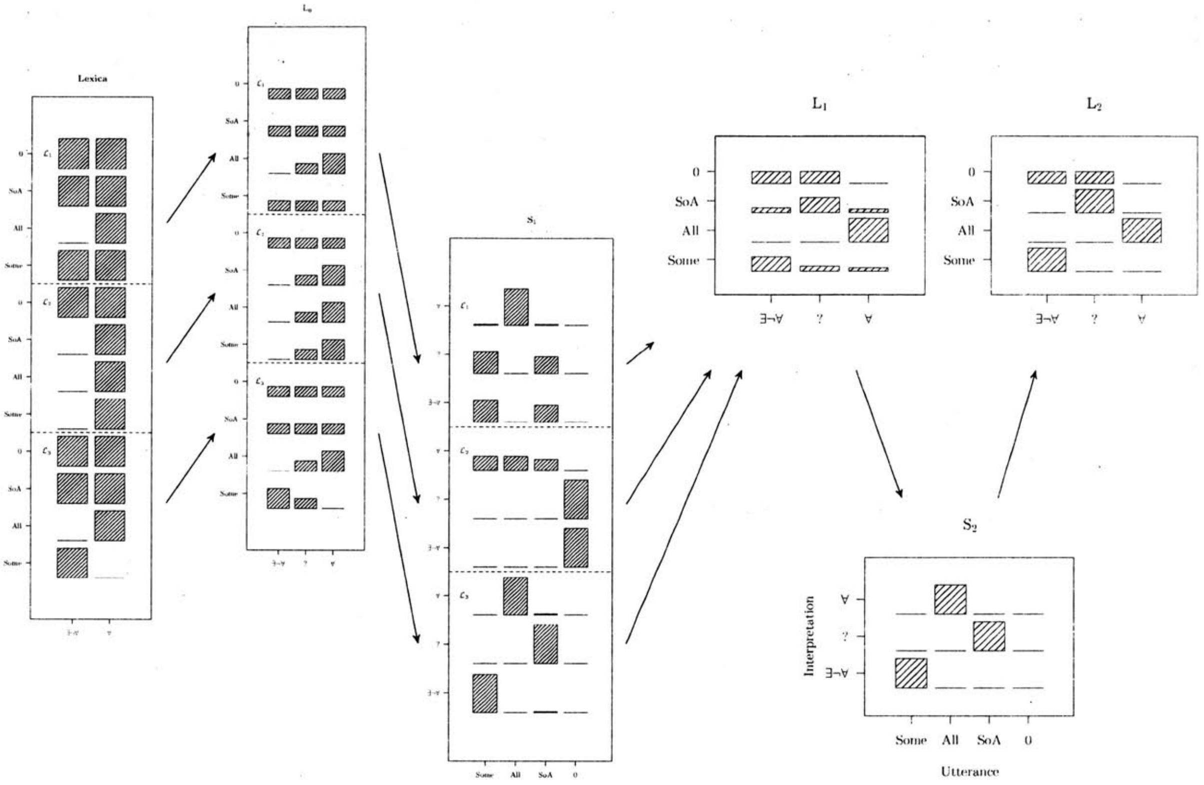


Figure 2-9: *Some or all* ignorance implicature under compositional lexical uncertainty.

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 $\forall$.

2.4.6 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 [33, 46].

Consider the following example:

(8) John didn't talk to Mary or Sue.

Without embedding under negation, as in (9) below, the disjunction would license a pragmatically strengthened, exclusive-or (XOR), meaning:

(9) John talked to Mary or Sue.

It has long been observed that when certain speaker knowledgeability assumptions are in the common ground, (9) 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 (8), 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 (8), 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 (8) 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 (8). 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 (8) in our framework. We assume that there are four worlds, $\mathcal{W} = \{\{\}, \{\text{Mary}\}, \{\text{Sue}\}, \{\text{Mary, Sue}\}\}$, where 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 O is maximal with respect to the worlds, i.e. 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, with $[\![m]\!] = \{\{\text{Mary}\}, \{\text{Mary, Sue}\}\}$, and s, with $[\![s]\!] = \{\{\text{Sue}\}, \{\text{Mary, Sue}\}\}$), 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 2-10 shows the nine possible refined lexica for this model, and the ultimate predicted pragmatic interpretation of Example (8), 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 interprets $\neg(m \vee s)$ as $\{\{\}\}$ with probability 1—thus, Example (8) 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

[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 2.4, 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 2-10. 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 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 knowledges state, and will therefore block the association of $\neg(m \wedge s)$ with this knowledge state.

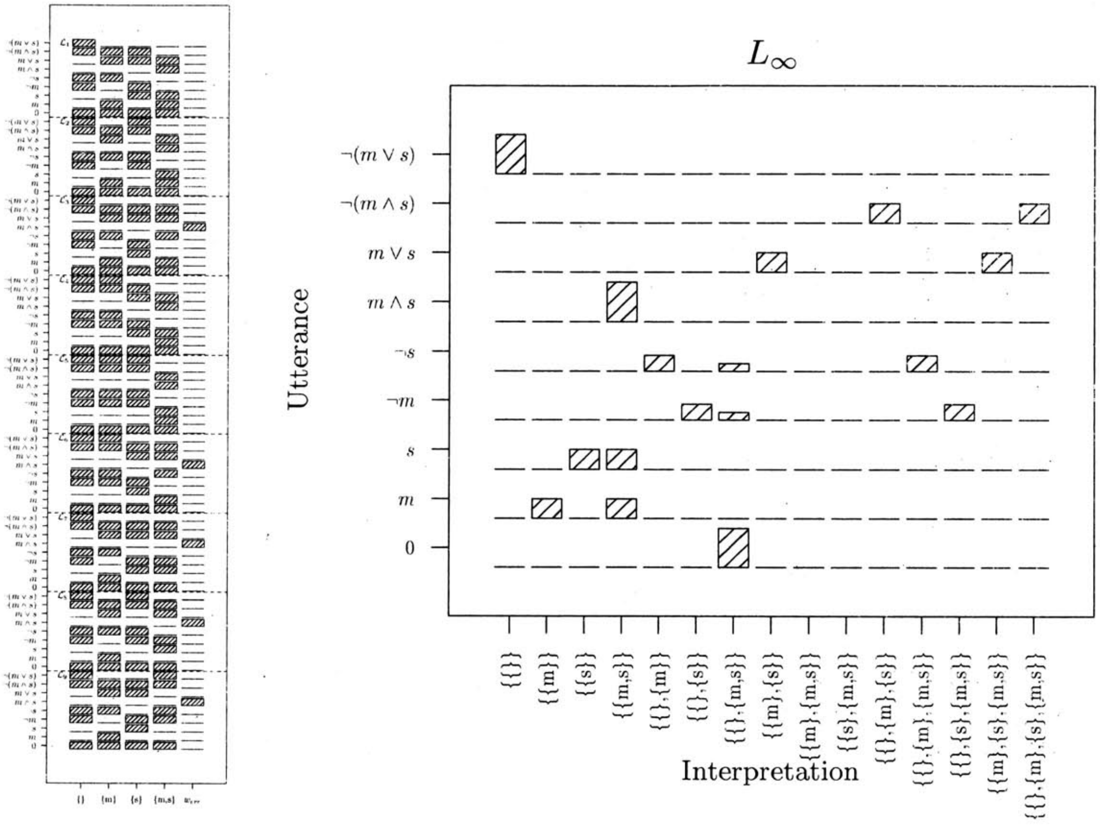


Figure 2-10: The set of lexica (left) in the negation/disjunction example (8), and the interpretation that the listener L_∞ assigns to each alternative utterance (right). The y-axis shows the 9 alternative utterances, while the x-axis shows the 4 possible worlds for the lexica, and the 15 possible knowledge 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.

explanation. First, we will explain why a speaker who wants to communicate this world will choose Example (8). 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 (8), 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, i.e. 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 2-10, 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 2-10:

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

We will give the explicit logic for why $\neg(m \vee s)$ is not the preferred utterance to express the critical XOR knowledge state $\{\{\}, \{\text{Mary, Sue}\}\}$; 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 2-10, $\neg(m \vee s)$ is compatible with the XOR knowledge state $\{\{\}, \{\text{Mary, Sue}\}\}$ 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 $\{\{\}, \{\text{Mary, Sue}\}\}$, but rather $\neg m$ or $\neg s$.

The model implementation we have just described illustrates why embedded implicature for Example (8) 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 (8) 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 (8), 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 [46]:

(10) 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 if a proper treatment would derive the embedded implicature in Example (10), thus remain as additional open questions.

2.4.7 Exceptional downward entailing contexts

The examples discussed in Section 2.4.6 seem to provide evidence that, in the absence of accenting, embedded implicatures do not occur in downward-entailing contexts. Indeed, there are several theoretical proposals which have been developed to account for this generalization [16, 27]. [16], 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 (8), as the reading associated with the implicature is logically weaker than the attested reading. [27] propose an extension of this condition, which makes the further prediction that Hurford-violating disjunctions cannot be embedded in downward-entailing contexts.

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 2.4.1. Consider the following examples:

(11) 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 degrees F (30-35 degrees C), or between 105-115 degrees F (40-45 degrees C), 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 degrees F or between 95-105 degrees F.]

(12) 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 (11), the speaker embeds the Hurford-violating disjunction "warm or scalding" under negation. As discussed in Section 2.4.1, 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*, i.e. *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.

We have already shown that lexical uncertainty can explain the lack of embedded implicatures in Example (8). We will now show that it can simultaneously explain the embedded implicature generated in Example (11). It is the distinctive structure of the sets of worlds and alternatives in Example (11) which lead to this implicature. The conditions which prevented an embedded implicature from being generated in Example (8) 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 degrees F, world 2 to water between 85-95 degrees F, world 3 to water between

95-105 degrees F, and world 4 to water between 105-115 degrees 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 (11) is represented by $\neg(w \vee s)$ in this formalization.

Figure 2-11 shows the predicted pragmatic interpretation of each alternative utterance, when, as in Figure 2-10, 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 (8) as discussed in Section 2.4.6: 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 (8) and (11) in our compositional lexical uncertainty model? In our discussion the utterance in Example (8), we outlined the following logic: for an utterance u and knowledge state o, is it the case that (i) 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 (8), 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 (11), 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 2.4.6, a key component of the reasoning lies in considering the number of different refined lexica in which various literal interpretations are available. Figure A-2 in Appendix A.3 depicts the 21 refined lexica for this problem, though for

[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.

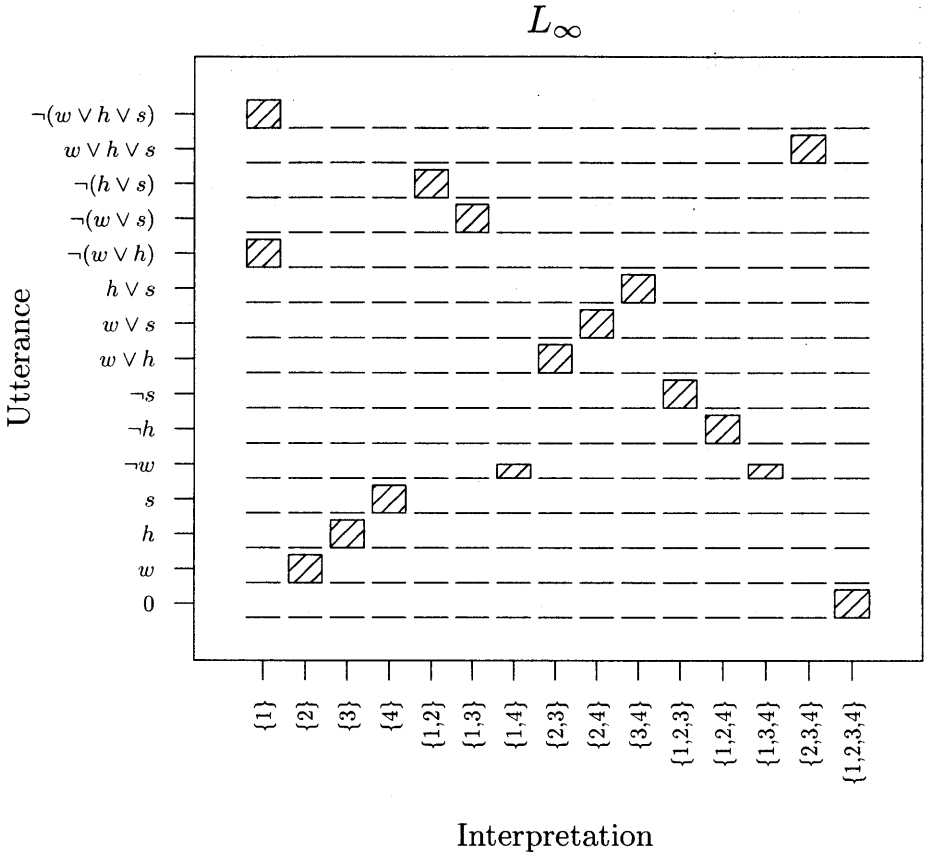


Figure 2-11: The interpretation the listener L_∞ assigns to each alternative utterance, in our formalization of Example (11). The y-axis shows the 15 alternative utterances, and the x-axis shows the 15 possible knowledge states. The cost of each disjunct is 0.1, the cost of negation is 0.1, the inverse-temperature λ is set to 5, and all knowledge states receive uniform prior probability.

discursive simplicity we will not refer directly to specific lexica in this section.

We elucidate this logic first by explaining part (i): why the speaker with knowledge state $\{1,3\}$ will use Example (11). 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 2.4.6. Unlike the case of Section 2.4.6, 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 2-11 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 2-11; 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 (11), though it does not predict this implicature in all cases. We briefly note that the implicature in Example (11) 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.

2.5 Conclusion

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 [36], 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, focussing 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 heterogenous effects within downward-entailing contexts: while many embedded implicatures are canceled within downward-entailing contexts, some with particular scale structure can survive.

Chapter 3

The strategic use of noise in pragmatic reasoning

3.1 Introduction

Prosody can be used in a highly productive manner to change the interpretation of utterances in natural language. There are a number of systematic changes in meaning which are associated with prosody. Consider the following sentence:

(1) JOHN went to the party.

(We use capital letters here to indicate that the speaker has placed stress on *John*; this convention is used throughout.) The placement of stress in this example leads to several inferences. First, it provides information about the speaker's communicative intentions, in particular about the question that the speaker is trying to answer. In this case, it is likely that the speaker wants to answer the question, *Who went to the party?* Second, it provides information about first-order facts about the world. In most contexts, the placement of stress on *John* will result in an exhaustive interpretation of the utterance. That is, it leads to the inference that nobody but John went to the party. Finally, the use of stress can provide information about the discourse context, in particular that the speaker is probably trying to correct a misconception on the part of the listener. In the current example, the speaker may be trying to communicate that the listener holds an incorrect belief about who went to the party, e.g. that Alice went rather than John.

These effects are highly general: the meaning of nearly any sentence in English can be systematically shifted by the use of stress. The effects of stress are also highly heterogenous, and the three which are discussed above do not exhaust their range. The interpretation of stress is sensitive to the background context of the conversation, the grammatical structure of the sentence used by the speaker, and the semantic content of this sentence. Stress can often strengthen the figurative use of utterances:

(2) Richard is an ANIMAL.

(3) She owes me A MILLION dollars.

(4) Yeah, I'd LOVE to edit your thesis for you.

The use of stress in these examples reinforces that the speaker intends them to be interpreted figuratively (as metaphor, hyperbole, and sarcasm, respectively). These cases also fall broadly into the category of *emphatic* uses of stress. The placement of stress on *animal* in Example (2) indicates that Richard is especially animal-like along the dimensions which are relevant in the context, and that the speaker has especially strong feelings about this, e.g. of admiration, disgust, etc.

A third class of effects are tied to specific lexical items used in an utterance. In the following cases, the placement of stress shifts the interpretation of *only* and *even*:

(5) Mary only introduced BOB to Alice.

(6) Mary only introduced Bob to ALICE.

(7) The police even ESCORTED the prince to the airport.

(8) The police even escorted the PRINCE to the airport.

These effects are typically labeled by the term *association with focus*. In Examples (5) and (6), the interpretation of the lexical item *only* changes depending on whether stress has been placed on *Bob* or *Alice*. If stress is placed on *Bob*, then the utterance indicates that nobody but Bob was introduced to Alice. In contrast, if it is placed on *Alice* then the utterance indicates that Bob was introduced to nobody besides Alice. A similar shift in interpretation occurs for *even* in Examples (7) and (8). When stress is placed on *escorted*, the utterance indicates that the police did other things for the prince besides escorting him; when it is placed on *prince*, the utterance indicates that the police escorted others besides the prince.

This work will be studying these three classes of phenomena, along with several others. While we will be presenting models of these phenomena, we do not intend these models to be interpreted as a theory of stress interpretation *per se*. Rather, we will be using these models to illustrate a set of pragmatic principles which can be used to explain these phenomena. These principles are the primary object of study in this work: while we do not propose that people use our particular models in order to derive the interpretation of stress, we do propose that they use these more abstract principles (in concert, most likely, with a number of others which have not yet been identified). These principles serve a second function, in helping to understand the structure of different stress-based phenomena. Different principles are required in order to derive different phenomena, and thus the success of certain principles in deriving a particular phenomenon reveals information about the structure of that phenomenon.

3.1.1 Stress and focus

Nearly all contemporary theories of stress interpretation make use of the notion of *focus* [84, 99, 12, 58, 5, 80]. Under these accounts, the placement of stress within a sentence is used to indicate which part of the sentence is focused. While stress refers to certain acoustic properties of a sentence's pronunciation, focus refers to an abstract property of a sentence's syntactic/semantic representation. Consider the following sentence:

(9) Mary fishes for BASS.

The claim that the word *bass* received stress is a claim about the acoustic properties of this sentence. In particular, it is the claim that the loudness and duration of the word *bass* were increased relative to the rest of the utterance, that its pitch was changed, etc.

The focus of the utterance, in contrast, is meant to describe the part of the utterance that is most prominent to the listener. In Example (9), if the placement of stress results in *bass* being the most prominent part of the utterance, then this would be indicated by the placement of focus on the word:

(10) Mary fishes for $[\text{bass}]_F$.

The notation $[\text{bass}]_F$ indicates that *bass* has been *focus-marked.* The placement of focus on *bass* clearest in this example if the utterance has been used to answer a question. For example, suppose that the following question had been previously asked in the discourse:

(11) What does Mary fish for?

If the speaker is responding to this question, then it is common knowledge that the speaker is trying to communicate what Mary fishes for. Intuitively, the most prominent part of the sentence will be the answer to this question, namely *bass.* The placement of stress in Example (9) thus would correspond directly to the placement of focus.

This direct correspondence between stress and focus does not, however, hold in all cases. Consider a dialogue involving a different question:

(12) A: What does Mary do for fun?
B: Mary fishes for BASS.

Here, the placement of stress on *bass* remains the same, but the speaker is providing a different answer through their utterance: they are communicating that what Mary does for fun is fish for bass. Thus, the focus of the utterance is broader than before, attaching to the entire phrase *fishes for bass*:

(13) Mary $[\text{fishes for bass}]_F$.

Focus therefore exists at a higher level of abstraction than stress, and cannot be directly read off of the placement of stress, or more generally the acoustic properties of the utterance.

The precise relationship between stress and focus is a complex; there are a number of proposals for how stress placement generates focus placement, and how the placement of stress constrains the placement of focus [89, 50, 39, 82].

Under standard accounts, stress does not have any direct effect on the interpretation of an utterance; the placement of focus mediates all of the interpretive effects of stress. There are a number of theories of how the placement of focus changes utterance interpretation. Here, we will review one of the best known, Alternative Semantics [83, 84], and in the process illustrate some high-level properties which are common to most of the major accounts.

According to Alternative Semantics, focus triggers a set of alternatives to what the speaker said. For example, consider again the utterance with focus on *bass*:

(14) Mary fishes for $[\text{bass}]_F$.

Under this account, the placement of focus on *bass* makes the listener consider a set of alternative meanings to *Mary fishes for bass.* These alternative meanings can, to a first approximation, be generated by replacing *bass* with other words: *Mary fishes for trout*, *Mary fishes for carp*, etc. More formally, this set of alternative meanings can be described by a set of propositions (or, under other formulations, a set of properties):

$$\{fishes(m,x)|fish(x)\} \tag{3.1}$$

Thus, when focus is placed on *bass*, its first effect is to make the listener consider other propositions of the form *Mary fishes for x.* Alternative Semantics provides a fully compositional theory of how to generate the set of alternatives for an utterance given focus placement, though we will not review the details of this theory here [73, 84].

Once a set of alternatives has been generated from focus, Alternative Semantics provides an account of how this set of alternatives will be used for interpretation. While the details are somewhat subtle, the broad outline can be illustrated by the current example. The theory states that the possible answers to the question under discussion must be a subset of the set of alternatives. Certain questions under discussion will satisfy this constraint, while others will not. Consider first the question *What does Mary fish for?* The possible answers to this question will be of the form *Mary fishes for x*, and every answer of this form will be a member of the alternatives set in Equation 3.1. Thus, the question *What does Mary fish for?* satisfies the constraint imposed by Alternative Semantics. In contrast, consider the question *Who fishes for bass?* In this case, the possible answers to this question will be of the form *x fishes for bass*. These possible answers are not members of the alternatives set in Equation 3.1, and as a result *Who fishes for bass?* is excluded as the question under discussion.

Alternative Semantics thus provides an account of how the question under discussion is constrained by the placement of focus. These constraints translate fairly directly into effects on the listener's interpretation of the utterance. If the listener knows that the speaker is answering the question *What does Mary fish for?*, then they can infer that the speaker is, to the best of their ability, providing a complete answer to this question. Thus, *Mary fishes for bass* will be interpreted as meaning that Mary does not fish for anything besides bass. In general, this mechanism allows focus to generate *exhaustive* interpretations, i.e. the inference that the speaker has provided an exhaustive answer to the inferred question.

3.1.2 The noisy-channel proposal

Focus is a full-fledged part of a language's semantic/syntactic representation under Alternative Semantics and related proposals. As already discussed, focus is related to, but not straightforwardly reducible to, the placement of stress within a sentence. Claims about focus within these theories therefore cannot be recast as claims about stress. Moreover, focus is used to determine the set of alternatives under these accounts, which in turn drives the semantic effects which these theories aim to explain. Focus plays a critical, and not easily eliminated, role in these theories.

This work will present a theory of stress interpretation without focus. Under our proposal, stress has no effect on the semantic/syntactic representation of an utterance. All of the interpretive effects of stress will be derived from its acoustic properties, and the effects that these properties have on pragmatic reasoning.[1]

[1] The pragmatic effects which we are interested in are not tied to acoustics *per se*. Indeed, it is possi-

In the absence of focus, or any semantic/syntactic representation of stress, it is also less natural to posit conventionalized rules for interpreting stress. Thus, in contrast to Alternative Semantics or other previous proposals, our account does not have any rules of the form: *The stress/focus of an utterance is used to generate alternative utterances according to the following algorithm...*, or *Focus-sensitive adverbs such as "only" use stress/focus to determine their interpretation using the following rules....* The interpretation of stress will be derived solely from general-purpose pragmatic reasoning.

We propose that the use of stress is viewed as the intentional decision to reduce the noise rate on the selected portion of the utterance. There are three main acoustic changes associated with prosodic stress: increased loudness, duration, and changes to the fundamental frequency [11]. An utterance that is louder and longer is less likely to get swamped by sounds in the environment, while changes in pitch will focus the listener's attention on the utterance. Thus, by placing stress on part of an utterance, the speaker is choosing to decrease the noise rate on that part of the utterance. The speaker's choice to reduce the noise rate is an intentional action, and thus will receive a pragmatic interpretation by the listener. The listener considers which intended meanings would have motivated the speaker to reduced the noise rate on the chosen part of the utterance, and infers that stress is more likely to indicate such meanings. The interpretive effects of stress are thus derived as pragmatic effects of the decision to reduce the noise rate.

3.2 Inference in a noisy-channel model

Communication channels are limited in various ways, and successful communication often requires complex inferences in order to overcome these limitations [90]. When a speaker in a conversation tries to communicate something to the listener, their intended signal may be corrupted by speech errors (e.g., if they are excited or intoxicated), environmental noise (if they are in a loud setting, e.g. at a cocktail party), or perceptual noise (if the listener is not paying attention). For the listener to successfully understand the speaker's intended

ble to derive the same effects for written communication, when stress is signaled by boldface, italics, or capitalization.

meaning, they must take into account these possible sources of noise, and infer whether they heard what the speaker actually intended. A growing body of experimental evidence suggests that people account for the possibility of noise when interpreting language; this has been successfully modeled as probabilistic inference of the original message given the received message [63, 64, 34, 8, 2, 19, 23].

In our pragmatics model, we will assume that the listener is aware of the possibility of noise, and rationally accounts for this possibility when interpreting utterances. Our model will draw on previous work which demonstrates how to optimally account for the possibility of noise during interpretation. Suppose that the speaker intends to send utterance u_i to the listener, and chooses prosodic stress s. The distribution over utterances which are *perceived* by the listener is given by $P_N(\cdot|u_i,s)$, where P_N is the *noise distribution*. In particular, the expression $P_N(u_p|u_i,s)$ is the probability that the listener will perceive utterance u_p given that the speaker intended utterance u_i and chose prosodic stress s.

When the listener hears an utterance u_i, they have uncertainty about what utterance was intended by the speaker. In general, there will be two explanations for why they perceived this utterance: either there was no noise, and the speaker's intended utterance was faithfully sent, or there *was* noise, and the speaker actually intended to send some other utterance. We assume that the listener has access to the noise distribution P_N. Therefore, in order to recover the utterance that was intended by the speaker, they can invert this noise process, by considering which intended utterances u_i would have been most likely to produce the utterance u_p which they perceived. This process of inverting the noise distribution P_N is defined by:

$$P(u_i|u_p,s) = \frac{P(u_i)P_N(u_p|u_i,s)}{\sum_{u_j} P(u_j)P_N(u_p|u_j,s)} \tag{3.2}$$

$$\propto P(u_i)P_N(u_p|u_i,s) \tag{3.3}$$

This equation defines the probability $P(u_i|u_p,s)$ that the speaker intended utterance u_i, given that the listener perceived utterance u_p with prosodic stress s. By Bayes' rule, this quantity is proportional to the product of two probabilities: the prior probability $P(u_i)$ that the speaker intended utterance u_i, and the probability $P_N(u_p|u_i,s)$ that the listener would perceive utterance u_p given that the speaker intended u_i and used stress s. Given a perceived utterance u_p, the utterance u_i is likely to have been intended if the speaker was *a priori* likely to have used this utterance, and if the perceived utterance was a probable corruption of this utterance.

3.3 Integrating RSA and Noisy-Channel Models

We will demonstrate that novel pragmatic inferences can be driven by the physical properties of the communication channel. Consider that agents in a conversation are typically mutually aware of the physical limitations of their communication system, e.g. that their utterances may be corrupted by noise. These agents can therefore take actions that exploit these limitations, and their conversational partners' knowledge of these limitations. Although noise is typically viewed as a problem for communication, we will argue that the *possibility* of noise is in many circumstances a resource, without which certain types of communication would not be possible.

We will model of the effects of noise on pragmatic reasoning by integrating the noisy-channel inference model defined above with the Rational Speech Act model, defined in Section 2.1. The literal listener, as defined in the RSA model, interprets utterances according to their literal meaning. The process of literal interpretation is more complicated, however, when the listener is uncertain about which utterance the speaker intended. In order to deal with this uncertainty, we propose an extension of the RSA literal listener. The listener in this case first tries to infer what utterance the speaker intended, and then considers the literal interpretation of this inferred utterance. More precisely, for each world w, the listener first computes the probability $K(w|u_p,s)$ that w is compatible with the speaker's intended utterance:

$$K(w|u_p,s) = \sum_{u_i} P(u_i|u_p,s)\mathbb{1}_{w\in[\![u_i]\!]} \tag{3.4}$$

Here, $\mathbb{1}_{w\in[\![u_i]\!]}$ is an indicator variable, which equals 1 if $w \in [\![u_i]\!]$ and 0 otherwise. The listener considers all of the utterances u_i which may have been intended by the speaker. For each utterance u_i, the listener determines whether the literal meaning of the utterance is compatible with the world w. Then, the listener weights utterance u_i by $P(u_i|u_p,s)$, the probability that u_i was actually intended by the speaker. If an utterance is likely to have been intended by the speaker, then it will have higher weight, and will contribute more towards determining whether the world w is compatible with what the speaker intended. By summing over all utterances u_i, the listener computes the marginal probability that w is compatible with the speaker's intended utterance.

The probability that the literal listener assigns to world w given utterance u_p is defined by:

$$L_0(w|u_p,s) \propto P(w)K(w|u_p,s) \tag{3.5}$$

The term $P(w)$ is the prior probability that the listener assigns to world w. The probability that the listener assigns to w is proportional to the prior probability of w, and the probability that the speaker's intended utterance is compatible with w. Worlds with higher prior probability, and which are more likely to be compatible with the speaker's intended utterance, are assigned higher probability by the listener.

The speaker is defined in a manner which generalizes the RSA definition along two dimensions. First, we assume that the speaker does not necessarily want to communicate the exact world state that they are in. Rather, they may have a question-under-discussion (QUD) which they want to answer. The QUD determines a partition on world states, and the speaker wants to communicate the cell in this partition which their world state belongs to [38]. Given a world state w and a QUD q, we first compute $R_n(u_p,s|w,q)$: the probability that the listener L_{n-1} assigns to the intended cell of the QUD partition, given that they perceive utterance u_p with prosodic stress s:

$$R_n(u_p,s|w,q) = \sum_{w'} \mathbb{1}_{q(w')=q(w)} L_n(w'|u_p,s) \tag{3.6}$$

The indicator variable $\mathbb{1}_{q(w')=q(w)}$ equals 1 when the world w' is in the same partition cell as the actual world w, and 0 otherwise. We assume that the QUD q is a function, which maps each world to its corresponding partition cell. The equation for R_n sums over all of the worlds which are in the same partition cell as w, weighing each world by the probability that it is assigned by the listener L_n.

The *information utility* that the speaker receives if the listener perceives utterance u_p (with prosodic stress s) is defined by:

$$I_n(u_p, s|w, q) = -\log \frac{1}{R_n(u_p, s|w, q)} \tag{3.7}$$

The quantity $\log \frac{1}{R_n(u_p, s|w, q)}$ is the surprisal that the listener assigns to the speaker's intended QUD cell. In the RSA model, the speaker wants to minimize the surprisal that the listener assigns to the intended world state. In the current model, the speaker wants to minimize the surprisal that the listener assigns to the intended QUD cell.

These definitions generalize the RSA model to communication about the QUD, rather than about the world. In order to model the current situation, we also need to generalize the model along another dimension. When the speaker chooses an utterance u_i, they do not know which utterance the listener will actually perceive. In order to compute the value of an utterance, they therefore need to consider the distribution over perceived utterances. We assume that the speaker has access to the noise model P_N, and therefore that they can compute this distribution. The *expected information utility* of the utterance u_i with prosodic stress s is defined as:

$$\mathbb{E}_{P_N(\cdot|u_i,s)} I_n(\cdot|w, q) = \sum_{u_p} P_N(u_p|u_i, s) I_n(u_i, s|w, q) \tag{3.8}$$

For each utterance u_p which the listener may perceive, the speaker first calculates the information utility that will be gained if the listener does perceive this utterance. The speaker then weighs each of these utilities by the probability $P_N(u_p|u_i, s)$ that the listener will perceive utterance u_p, and sums across these possibilities to compute the average information utility associated with intended utterance u_i.

The speaker S_n chooses an utterance and a prosodic stress in order to maximize their expected information utility, while simultaneously minimizing their cost. The speaker's utility function is defined by:

$$U_n(u_i, s|w, q) = \mathbb{E}_{P_N(\cdot|u_i,s)} I_{n-1}(\cdot|w, q) - c(u_i) - c(s) \tag{3.9}$$

Here $c(u_i)$ is the cost of utterance u_i and $c(s)$ is the cost of prosodic stress s. Note that the speaker S_n considers the listener L_{n-1}'s interpretation of their utterance when computing the expected information utility. As in the normal RSA model, the speaker uses a softmax decision rule to choose utterances:

$$S_n(u_i, s|w, q) \propto e^{\lambda U_n(u_i, s|w, q)} \tag{3.10}$$

The pragmatic listener L_n ($n > 0$) interprets a perceived utterance u_p by integrating two types of information. For each world w and QUD q, the listener first considers how likely the speaker S_n would have been to choose a particular intended utterance u_i, given this world and QUD. Then, the listener considers how likely this intended utterance would have been to produce the perceived utterance u_p, given the noise distribution P_N. A particular world and QUD are inferred to be likely, if they are likely to have produced the perceived utterance u_p. The probability $L_n(w, q|u_p, s)$ that the listener assigns to world w and QUD q, given that utterance u_p and prosodic stress s were perceived, is defined as follows:

$$L_n(w, q|u_p, s) \propto P(w)P(q|w) \sum_{u_i} S_n(u_i, s|w, q) P_N(u_p|u_i, s) \tag{3.11}$$

The term $P(q|w)$ is the conditional probability that the speaker's QUD is q, given that the world is w.

3.4 Relationship between prosody and QUD

In this section, we will explain a primary factor which drives the interpretation of prosodic stress in this model: inferences about the speaker's QUD. When the speaker places prosodic stress on part of their utterance, they are making an intentional decision to reduce the noise rate on that part of the utterance, at a certain cost. They will only make this decision when the benefits from reducing the noise rate — and increasing the probability that the listener will accurately perceive the selected portion of the utterance — exceed the cost of using prosodic stress.[2] Depending on the speaker's QUD, it may or may not be rational to use prosodic stress whatsoever, and the optimal placement of stress within the sentence may differ as well. The speaker will place prosodic stress on the parts of the utterance that are especially important for accurately communicating the answer to their QUD. These are the parts of the utterance, such that if the listener mishears these portions, they will infer an incorrect answer to the speaker's QUD.

The listener knows all of this, and tries to find an explanation of the speaker's choice to place stress on a portion of the utterance. The listener knows that the speaker only uses prosodic stress when doing so is important for accurately communicating the answer to the speaker's QUD. The listener therefore tries to identify the QUDs which would rationalize this decision, i.e. which would make the selected portions of the utterance especially important for answering the QUD. Certain QUDs *will* rationalize this decision, and certain QUDs will not. The listener will place higher probability on those QUDs which provide an adequate explanation of the speaker's behavior.

In order to illustrate this reasoning, we will start with a minimal communication game, which includes a highly simplified representation of prosodic stress. This game is intended as a formalization of the contrast between the following examples:

(15) Bob went to the store.

(16) BOB went to the store.

We suppose that there are two people, Alice and Bob, who may have gone to the store, either individually or together. These possibilities are represented as three possible worlds: $\{Alice\}, \{Bob\}, \{Alice, Bob\}$. The speaker knows exactly who went to the store, while the listener has a uniform prior distribution over worlds.

There are two different types of speaker QUDs in this example: *polar-QUDs* and *list-QUDs*. A polar-QUD is a partition with two cells, i.e. a yes/no question. In this case, it is a question of the form *Did X go to the store?* where *X* is either *Alice* or *Bob*. We denote

[2] This is not strictly speaking true, as the speaker's softmax decision rule will sometimes lead them to make suboptimal choices. However, the speaker is still strictly more likely to choose an action with higher utility than one with lower utility, and this informal reasoning will generally lead to valid conclusions about the model's behavior.

these QUDs by q_A and q_B. A list-QUD is a question which asks for an answer in the form of an exhaustive list.[3] In this example, it is the question *Exactly who went to the store?* We denote this QUD by q_L. For convenience, both types of QUDs are represented as functions: the polar-QUDs are represented as functions from the set of worlds to $\{\top, \bot\}$, and the list-QUD is represented as a function from worlds to individuals (i.e. the value of the function on a world is the set of individuals who went to the store in that world).

The distribution over QUDs is defined by:

$$P(q_A|\{Alice\}) = 0.5; P(q_L|\{Alice\}) = 0.5$$
$$P(q_B|\{Bob\}) = 0.5; P(q_L|\{Bob\}) = 0.5$$
$$P(q_A|\{Alice, Bob\}) = \frac{1}{3}; P(q_B|\{Alice, Bob\}) = \frac{1}{3}; P(q_L|\{Alice, Bob\}) = \frac{1}{3}$$

In other words, if the speaker knows that Alice and only Alice went to the store, then they are equally likely to want to communicate the following two propositions: that (at least) Alice went to the store; and that only Alice went to the store. The situation is symmetric for the speaker who knows that Bob and only Bob went. For the speaker who knows that both Alice and Bob went, there is a uniform distribution over whether they want to communicate that (at least) Alice went, that (at least) Bob went, or that both Alice and Bob went.

The speaker has three alternative utterances available: a, b, and $a \wedge b$, with their semantics given as follows: $[\![a]\!] = \{\{Alice\}, \{Alice, Bob\}\}$, $[\![b]\!] = \{\{Bob\}, \{Alice, Bob\}\}$, $[\![a \wedge b]\!] = \{\{Alice, Bob\}\}$. We assume that the atomic utterances a and b have cost 1, while the conjunction $a \wedge b$ has cost 2. The qualitative effects in this section are robust to alternative cost values, given the constraint that $c(a) = c(b) < c(a \wedge b)$.

The prosodic stress term s is assumed to be binary: $s \in \{0, 1\}$, where $s = 1$ indicates that prosodic stress was used. If the speaker uses prosodic stress, then the probability of noise is decreased by a factor of 2. There is cost 0 for not using prosodic stress, and cost 0.1

[3] This definition is clearly very informal, but there will usually be an obvious way to make it precise in the examples which we consider.

for using it. We assume that the noise distribution P_N has an especially simple structure:

$$\begin{aligned}
&P_N(a|a,s) = 1-(p-\frac{s\cdot p}{2}); P_N(b|a,s) = (p-\frac{s\cdot p}{2})\\
&P_N(b|b,s) = 1-(p-\frac{s\cdot p}{2}); P_N(a|b,s) = (p-\frac{s\cdot p}{2})\\
&P_N(a\wedge b|a\wedge b) = 1
\end{aligned}$$

This noise distribution can be illustrated by Figure 3-1, for the case that prosodic stress is not used. The figure shows that if the speaker intends to choose utterance a, then this message will be accurately transmitted with probability $1-p$, and it will be corrupted to utterance b with probability p. The situation for utterance b is symmetric. In this simplified model, utterance $a\wedge b$ is always assumed to be transmitted accurately. The quantity p is the *noise probability*, and it is set to 0.01 in this example. If prosodic stress is used (and $s=1$), then the noise probability decreases from p to $p-\frac{s\cdot p}{2}=\frac{p}{2}$. The effects described in this section are quite robust to variations in the noise model, to variation in the magnitude of the noise-reduction effect from prosodic stress, and to changes to the cost of stress.

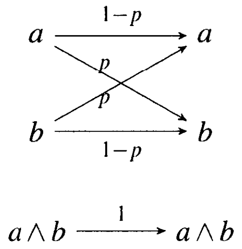


Figure 3-1: The noisy channel on the three utterances, when prosodic stress is not used.

There are two more parts of the model which need to be specified. First, in Equation 3.2, the literal listener uses a prior distribution over utterances in order to determine the speaker's intended utterance. This prior distribution is assumed to be uniform over utterances. (In fact, the prior probability of utterance $a\wedge b$ does not in fact influence the model's predictions, due to the structure of the noise distribution P_N. All that is required is that $P(a)=P(b)$.) Finally, the inverse-temperature parameter λ is set to 5.

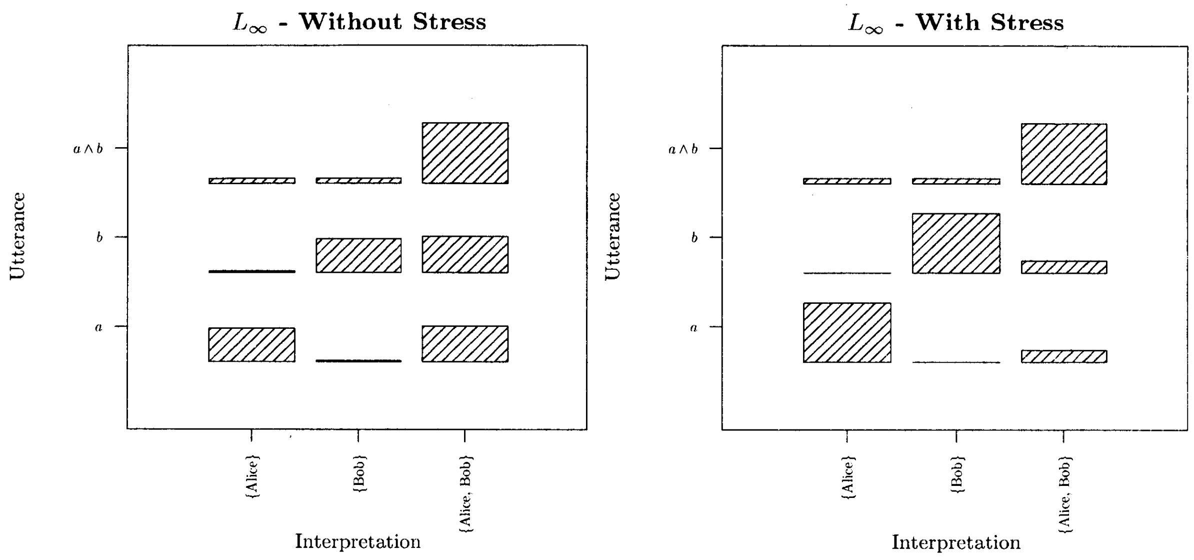


Figure 3-2: The listener's interpretation of the utterances without stress (on the left panel) and with stress.

3.4.1 Asymmetry between polar-QUDs and list-QUDs

Figure 3-2 shows the model's predicted listener distribution over worlds, given each utterance-prosody pair. The model predicts that the listener will interpret prosodic stress as indicating an exhaustive answer to the list-QUD. When the speaker places stress on utterance a, this is interpreted as meaning that only Alice went to the store, and similarly when the speaker places stress on b.

Consider a speaker in world $\{Alice\}$, i.e who knows that only Alice went to the store. This speaker may have one of two QUDs, the polar-QUD q_A or the list-QUD q_L. Regardless of which QUD they want to answer, this speaker will typically choose utterance a, which states that Alice went to the store. This is the simplest and most informative utterance for either QUD.

When the speaker chooses utterance a, they do not know which utterance the listener will perceive. As illustrated in Figure 3-1, if the speaker intends utterance a, then there is some probability that the listener will perceive utterance b instead. Therefore, when modeling how the listener will interpret their utterance, the speaker must take into account the possibility that the listener will perceive this utterance b, and not a. The literal meaning of utterance b is that Bob went to the store, so if if the listener perceives this utterance, then they will form very different beliefs than if they heard a.

The speaker can reduce the probability of noise, by placing prosodic stress on their utterance. In this case, if the speaker thinks that it is especially important for the listener to perceive the intended utterance a, rather than utterance b, then they can use prosodic stress to reduce the probability of misperception. As we will now explain, if the listener misperceives the utterance a, then this is more damaging for the speaker with the list-QUD q_L than for the speaker with polar-QUD q_A. As a result, it is more valuable for the speaker with the QUD q_L to reduce the noise rate than for the speaker with the QUD q_A. The speaker with QUD q_L will be more likely to use prosodic stress, and the listener, who knows this, will interpret prosodic stress as indicating this QUD.

The reasoning which drives the asymmetry between QUD types is quite general, and holds independent of particular model parameterizations. First, consider the speaker who wants to communicate the answer to the polar-QUD q_A, *Did Alice go to the store?* In the world $\{Alice\}$, the correct answer to this question is *yes*, so the speaker wants the listener to believe that (at least) Alice went to the store. Importantly, the speaker can correctly communicate the answer to this QUD, even if the listener has incorrect beliefs about other aspects of the world, in particular about whether Bob went to the store. This has the consequence that if the listener misperceives the speaker's intended utterance, the speaker may still have communicated the correct answer to their QUD. Consider a scenario in which the speaker has intended to use utterance a (which most directly communicates the answer to their QUD), and the listener misperceives this as utterance b. The speaker will still gain a substantial amount of utility in this case. There are two utterances which are compatible with the literal meaning of the utterance b: $\{Bob\}$ and $\{Alice, Bob\}$. In the latter world, the answer to the QUD is the same as in the actual world. The listener's beliefs are not so far from the meaning that the speaker wanted to communicate — the listener still assigns relatively high probability to the correct answer to the QUD. It is therefore not disastrous for the speaker if the listener mishears the utterance a as utterance b, and the speaker will not have a large incentive to decrease the noise rate through the use of prosodic stress.

This contrasts with the situation for the speaker with the list-QUD q_L. This speaker wants to communicate the answer to the question, *Exactly who went to the store?* In the world $\{Alice\}$, the correct answer to this question is *Alice*, so the speaker wants the listener to believe that Alice (and nobody else) went to the store. The only world in which this is true is $\{Alice\}$ — in every other world in the current example, someone else besides Alice went to the store. If the listener believes that the true world is any of these other worlds, then the speaker will not have communicated the correct answer to the QUD. The speaker will use utterance a in this situation, as every other utterance will communicate a false answer to the QUD. If the listener mishears this intended utterance, and instead perceives utterance b, then the speaker receives very low utility. The listener who hears utterance b will have false beliefs about the answer to the QUD — they will believe that someone else besides Alice went to the store. The speaker will therefore have a large incentive to prevent this from happening, and will use prosodic stress to decrease the probability of misperception.

The asymmetry between the list-QUD and the polar-QUD thus arises from an asymmetry in the consequences of noise for the two QUD types. For the speaker with the list-QUD, noise is disastrous: if the listener mishears the intended utterance, then the answer to the QUD will not be communicated correctly. For the speaker with the polar-QUD, noise is not disastrous: the correct answer to the QUD may still be communicated (with relatively high probability), even if the intended utterance is misperceived.

We have so far explained why the speaker will be more likely to place stress on their utterance if they have the list-QUD, and hence why the listener will be likely to interpret stress as indicating the list-QUD. The model makes a further prediction, shown in Figure 3-2: if the speaker places stress on the simple utterance a (or, by symmetry, on b), then the utterance is more likely to be assigned an exhaustified interpretation by the listener. That is, the listener is more likely to interpret the utterance as world $\{Alice\}$, in which only Alice went to the store, rather than world $\{Alice, Bob\}$, in which both Alice and Bob went to the store. The explanation for this is closely linked to the listener's inferences about the speaker's QUD. While the speaker in world $\{Alice\}$ will use stress if they have the list-

QUD q_L, there is no QUD that will induce the speaker in world $\{Alice, Bob\}$ to use stress. As a result, when the listener hears stress on utterance a, they will infer that the speaker is not in world $\{Alice, Bob\}$.

We have already explained why the speaker in world $\{Alice\}$ with QUD q_L is likely to place stress on utterance a. We will now explain why the speaker in world $\{Alice, Bob\}$ will not use utterance a with stress. Suppose first that this speaker has the list-QUD q_L. The speaker then wants to communicate that both Alice and Bob went to the store. The utterance which communicates the maximal amount of information for this speaker is $a \wedge b$, as its literal meaning is that both Alice and Bob went to the store. However, this utterance is costly (by assumption), and the speaker will sometimes choose a cheaper utterance which still provides some information. Due to symmetry, the speaker will have no preference among utterances a and b: the literal meaning of a rules out the world in which only Bob went to the store, while the literal meaning of b rules out the world in which only Alice went to the store. If the speaker chooses utterance a, then they will be indifferent about whether the listener mistakenly hears utterance b instead (and similarly if they choose utterance b). As a result, the speaker will have no incentive to place stress on their utterance if they choose either a or b.

Next consider the speaker in world $\{Alice, Bob\}$ with QUD q_A. This speaker knows that both Alice and Bob went to the store, and wants to communicate that (at least) Alice went. Suppose that the speaker chooses utterance a, and this is misperceived as b by the listener. The literal meaning of b is still compatible with the true world and the correct answer to the QUD; b literally communicates that at least Bob went to the store, which is compatible with both Alice and Bob going. It is therefore not disastrous for this speaker if the listener mishears utterance a as b, and this speaker will not have a strong incentive to decrease the noise rate using prosodic stress.

The reasoning in this section is worked through in more detail in Appendix B.1. The appendix provides explicit calculations for the speaker and listener distributions discussed in this section. In addition, it provides a formal illustration of how the asymmetry between the list- and polar-QUDs arises from model.

3.5 Prosodic stress and compositional structure

So far, we have explained why the use of prosodic stress will provide information about the speaker's QUD in a certain simple example. A crucial simplifying assumption in that example was that prosodic stress could be represented as a binary choice: either the speaker placed stress on the utterance, or they did not. This assumption clearly does not hold in actual language use, and indeed, the choice of stress location typically carried information about the speaker's intended interpretation:

(17) BOB went to the store.

(18) Bob went to the STORE.

In Example (17), the speaker is most likely answering the question *Who went to the store?* and is conveying the information that only Bob went to the store. In Example (18), the speaker is most likely answering a different question — *Where did Bob go?* — and is conveying that Bob only went to the store. Thus, our account needs to be able to explain the relationship between stress location and interpretation.

In order to provide a model of this relationship, we will extend our previous example along several dimensions. Our goal will be to provide a minimal example which demonstrates the link between stress location and interpretation; the same principles which generate this link in the example will also apply in more complex scenarios.

3.5.1 Model assumptions

We suppose that there are two people, Alice and Bob, and two locations, the store and the restaurant. Each person may have gone to either or both of the locations, and these decisions were made independently. The predicate S (R) indicates the set of individuals who went to the store (respectively, restaurant). A world is specified by the set of individuals satisfying each predicate. In the shorthand that we use, the world $\{\{Alice\}_S, \{Bob\}_R\}$ denotes the world in which only Alice went to the store, and only Bob went to the restaurant. We assume that at least one individual went to at least one location. As a result, there are 15 possible worlds, corresponding to $4^2 - 1$ permissible ways of assigning individuals to the two predicates. The speaker is assumed to have complete knowledge of the world, while the listener has a uniform prior distribution over worlds.

There are again two types of speaker QUDs: polar-QUDs and list-QUDs. The polar-QUDs are questions of the form, *Did X go to Y?* We allow X to be filled in with either of the two names, *Alice* or *Bob*, and Y to be filled in with either of the locations, *store* or *restaurant*. We use the notation $q_{P(X,Y)}$ to denote each polar QUD. For example, $q_{P(A,S)}$ denotes the polar question, *Did Alice go to the store?* There are two types of list-QUDs. The *Who*-QUDs are questions of the form, *Who went to Y?* The variable Y may be filled in with either of the locations. The *Who*-QUDs are denoted by $q_{Who(Y)}$. For example, $q_{Who(S)}$ denotes the QUD, *Who went to the store?* The *Where*-QUDs are questions of the form, *Where did X go?* The variable X may be filled in with either of the individuals. The *Where*-QUDs are denoted by $q_{Where(X)}$. For example, $q_{Where(A)}$ denotes the question, *Where did Alice go?* The QUDs are given their obvious semantics.

Given a world w, the distribution over QUDs is uniform over those which satisfy the following two properties: a) if the QUD is a polar-QUD, then its correct answer in world w must be *true*; b) if the QUD is a list-QUD, its correct answer in w must not be the empty set. For example, consider the world $\{\{Alice\}_S, \{\}_R\}$, in which the only event that took place was Alice going to the store. In this case there are three QUDs which satisfy these properties: *Where did Alice go?*; *Who went to the store?*; and *Did Alice go to the store?*. For every other QUD, the correct answer is either *false*, as in the case of *Did Bob go to the store?*, or it is the empty set, as in the case of *Who went to the restaurant?* We exclude QUDs which do not satisfy these two properties, in order to limit the number of required alternative utterances. If these QUDs were to be included, then their correct answers could only be communicated through utterances containing negation. In the previous example, the correct answer to *Did Bob go to the store?* is *false*, so in order to effectively communicate this, the speaker would need to have the utterance "Bob did not go to the store" as an alternative. Excluding these QUDs and utterances containing negation do not affect the qualitative predictions of the model, but will simplify our explanations.

The speaker's alternative utterances are generated from the following grammar:

$$S \rightarrow P : L$$

$$P \rightarrow a, \;\; P \rightarrow b, \;\; P \rightarrow a \wedge b$$

$$L \rightarrow s, \;\; L \rightarrow r, \;\; L \rightarrow s \wedge r$$

There are nine utterances generated by this grammar. Each utterance consists of a person phrase (to the left of the colon), and a location phrase. The nonterminal P determines the person phrase. This represents the set of individuals that the utterance is referring to — either a (Alice), b (Bob), or $a \wedge b$ (Alice and Bob). The nonterminal L determines the location phrase. This represents the locations that are being referred to — either s (the store), r (the restaurant), or $s \wedge r$ (the store and the restaurant). These utterances are given the obvious semantics. For example, the utterance $a \wedge b : s$ has the literal meaning that Alice and Bob went to the store. The utterances are assigned cost in proportion to their length. We assume that the atomic terms a, b, s, and r each have cost 1, and the total cost of an utterance is computed by summing the number of atomic terms that it contains.

There are three possible choices of prosodic stress s: we assume $s \in \{\perp, Left, Right\}$. If $s = \perp$, then this indicates that no prosodic stress was used. *Left* (*Right*) indicates that the speaker has placed stress on the phrase to the left (right) of the colon. For example, if the speaker uses stress $Left$ with utterance $a : s$, then this represents the speaker placing stress on "Alice" in the utterance "Alice went to the store." If the speaker uses stress $Left$, then this decreases the probability of noise on the left phrase of the utterance by a factor of 2, and similarly if they use stress $Right$.

The noise distribution is illustrated in Figure 3-3. The figure shows the noise distribution for person phrases (on the left) and location phrases (on the right). As in the previous example, we assume that the atomic person terms (a, b) may be confused for each other, and similarly for the atomic location terms (s, r), but the conjunctions ($a \wedge b$, $s \wedge r$) are always perceived accurately. We assume that the noise is sampled independently for the person phrases and location phrases. For example, if the speaker intends utterance $a : s$, then there is probability p that a will be perceived as b, and probability p that s will be perceived as r. Hence there is probability p^2 that the listener will mistakenly perceive utterance $b : r$.

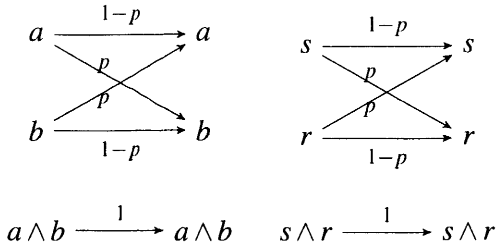


Figure 3-3: Each utterance is composed of a left phrase and a right phrase. The noise distribution for left phrases is shown on the left, while the noise distribution for right phrases is shown on the right. We assume that noise applies independently to the left and right phrases of the utterance. When prosodic stress *Left* is used, it decreases the noise probabilities in the left figure by a factor of 2, and similarly for prosodic stress *Right* and the right figure.

3.5.2 Interpretation of stress placement

Figures 3-4 and 3-5 show the interpretation of the different types of prosodic stress in this example. When the speaker uses prosodic stress, this indicates that they are trying to answer one of the list-QUDs. If they use stress *Left*, the listener infers that they are trying to answer a *Who*-QUD, and they interpret the utterance exhaustively with respect to this QUD. For example, if the speaker places stress on utterance $a : s$, the listener interprets this as meaning that nobody but Alice went to the store. In contrast, if the speaker uses stress *Right*, the listener infers that they are providing an exhaustive answer to a *Where*-QUD. If the speaker does not use prosodic stress, then this provides evidence that they are trying to answer one of the polar-QUDs. The results generalize those of Section 3.4. In that section, the lack of stress indicated a polar-QUD, while the use of stress indicated a list-QUD. In the current example, there are multiple types of list-QUDs, and multiple locations for placing stress. Different stress locations indicate different types of list-QUDs. We will address why these associations are generated by the model.

The intuition for why prosodic stress indicates a list-QUD remains the same as in the previous example: when the speaker has a list-QUD, rather than a polar-QUD, it is more important for the listener to perceive their utterance accurately, and as a result they will be more likely to use prosodic stress to reduce the noise rate. The crucial question in the current example is why prosodic stress on the left location indicates a *Who*-QUD, rather than a *Where*-QUD (and, by symmetry, why stress on the right location indicates a *Where*-

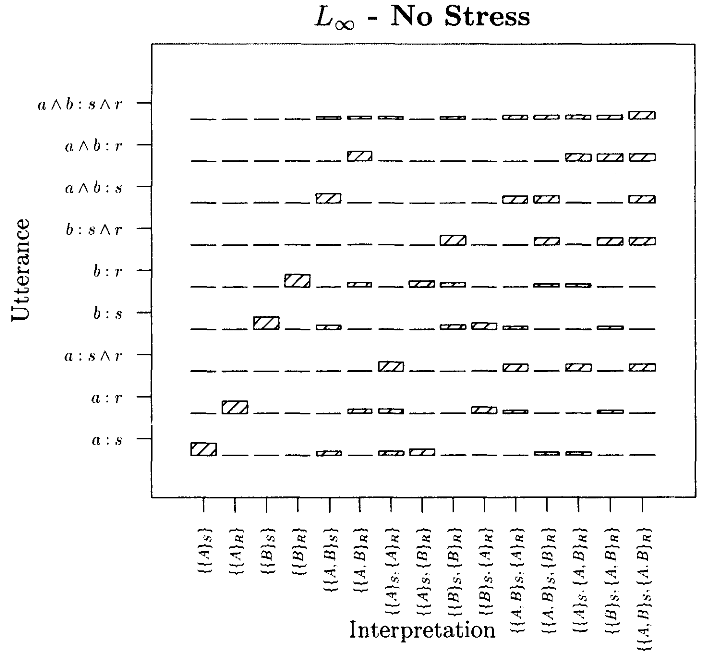


Figure 3-4: The listener's interpretation of the utterances without stress.

QUD, rather than a *Who*-QUD).

To be concrete, suppose that the speaker is in world $\{\{Alice\}_S, \{Bob\}_R\}$, i.e. the world in which only Alice went to the store, and only Bob went to the restaurant. Suppose further that this speaker has QUD $q_{Who(S)}$, i.e. *Who went to the store?* The most informative utterance for this speaker is clearly $a : s$, i.e. "Alice went to the store," and the speaker is therefore likely to choose this utterance. The speaker has two choices of prosodic stress: *Left*, in which case the noise rate on a will be reduced, and *Right*, in which case the noise rate on s will be reduced. We will explain why this speaker is more likely to reduce the noise rate on a rather than s, that is, why this speaker is likely to use stress *Left*.

When the speaker chooses utterance $a : s$, there are two types of noise that can occur (with reasonably high probability). First, it is possible that phrase a will be corrupted to phrase b, while phrase s will be communicated accurately. Second, it is possible that phrase s will be corrupted to phrase r, while phrase a will be communicated accurately.[4] If

[4] It is also possible that both a and s will be simultaneously corrupted. However, this case can be safely ignored when considering the model's predictions. Under our noise distribution, noise events across different parts of the sentence each occur independently with probability p. As a result, the probability of having two

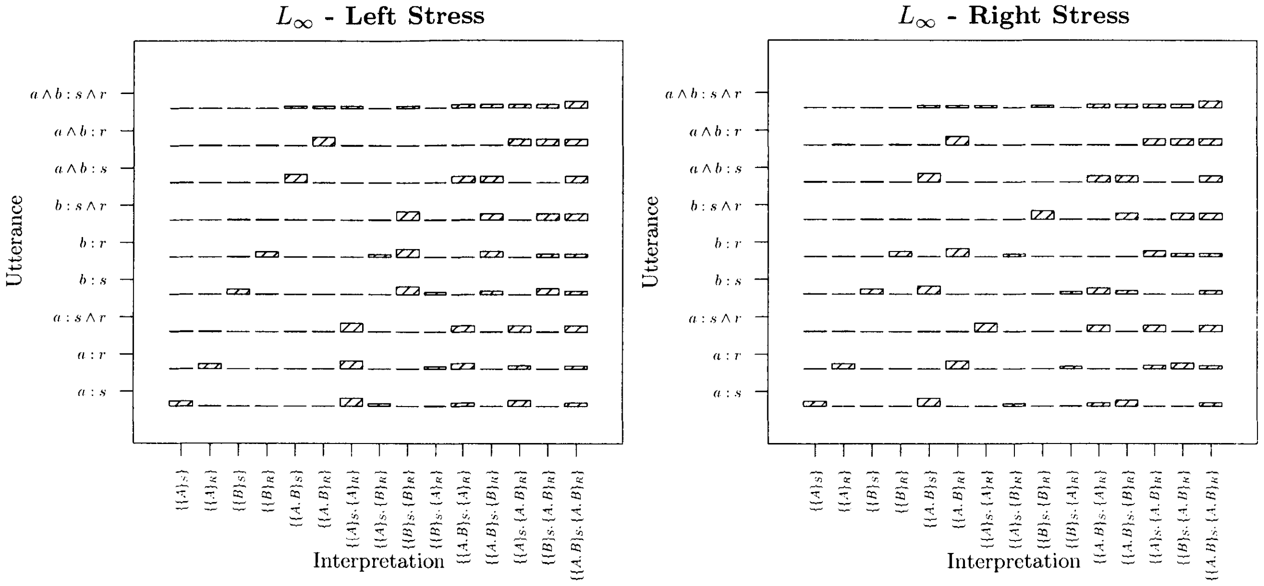


Figure 3-5: The listener's interpretation of the utterances with stress *Left* (on the left panel) and stress *Right*.

a is corrupted, then the listener will mistakenly believe that Bob went to the store. If *s* is corrupted, then the listener will mistakenly believe that Alice went to the restaurant. Which of these outcomes is worse for the speaker?

Given the QUD $q_{Who(S)}$, the speaker's goal is to communicate that Alice, and only Alice, went to the store. If the listener mishears *a* and forms the mistaken belief that Bob went to the store, this is disastrous for the speaker. The listener in this case will believe an incorrect answer to the speaker's QUD, as it is not possible in this case that only Alice went to the store. In contrast, if the listener mishears *s* and forms the mistaken belief that Alice went to the restaurant, this is not disastrous for the speaker. The listener's beliefs are still consistent with the correct answer to the QUD; even if Alice went to the restaurant, it is still possible that she (and only she) went to the store.[5] This produces an asymmetry between the different types of noise. The speaker with QUD $q_{Who(S)}$ will want to reduce the noise

simultaneous noise events is p^2, which is much smaller than p for plausible values of p. The speaker and listener will effectively ignore such low probability events in their reasoning.

[5] The fact that it is *a priori* possible that Alice went to both the store and the restaurant is of course contingent on our modeling assumptions. We will return below to the issue of whether the model's predictions are sensitive to world knowledge assumptions, and whether this sensitivity is empirically appropriate.

rate on the left part of the utterance more than they want to reduce the noise rate on the right part. As a result, they will be more likely to use the prosodic stress $Left$ rather than $Right$. The same general reasoning shows that the speaker with any *Who*-QUD will prefer to place prosodic stress on the name of the individual who is the answer to the question.

A symmetric result holds for the speaker with any *Where*-QUD. This speaker will prefer to use prosodic stress *Right* rather than $Left$, i.e. to place prosodic stress on the name of the location that is the answer to the question. For example, suppose that the speaker wants to answer the QUD $q_{Where(A)}$, *Where did Alice go?* Suppose further that Alice only went to the store, and that the listener uses utterance $a : s$ to communicate this. The speaker in this case will prefer to place prosodic stress on s rather than a.

The listener knows that the speaker is more likely to choose prosodic stress $Left$ given a *Who*-QUD, and stress *Right* given a *Where*-QUD. The listener also knows that the speaker is less likely to use prosodic stress of any sort to communicate a polar-QUD. As a result, when the speaker places stress on the left phrase in the utterance, the listener will infer that the speaker is more likely to have a *Who*-QUD. Similarly, the listener will associate stress on the right phrase with a *Where*-QUD.

3.6 Knowledgeability

Stress can be used to signal the precision of the speaker's communicative intent. In Sections 3.4 and 3.5, we explained why the use of stress will convey that the speaker has a more precise QUD, i.e. a list-QUD rather than a polar-QUD. As we will explain in the current section, this is not the only sense in which stress can signal precision. If there is *a priori* uncertainty about how much the speaker knows about a particular topic, then stress will often indicate that the speaker has a greater degree of knowledgeability. That is, stress will indicate that the speaker has a more precise knowledge state.

The interaction between stress and speaker knowledgeability can be illustrated using scalar implicatures [84]. Consider a scenario in which the speaker, a teacher, has graded (at least) some of the students' tests. The listener does not know whether all of them have been graded, and hence does not know whether the speaker knows how many of the students passed. Consider the following two utterances:

(19) Some of the students passed the test.

(20) SOME of the students passed the test.

In Example (19), the speaker does not place stress on the scalar item *some*. Because the speaker is not presumed to know whether all of the students passed, this utterance will typically generate a weak scalar implicature: that the speaker merely does not know that all of the students passed [46]. In contrast, when stress is placed on *some*, as in Example (20), the utterance will generate a strong implicature: that the speaker knows that not all of the students passed. The use of stress in this example thus signals a greater degree of knowledgeability for the speaker. Without stress, the utterance signals that the speaker does not know that the scalar alternative *all* is true. With stress, it signals that the speaker knows that the scalar alternative is false.

We will explain why the noisy-channel account predicts this interaction between stress and knowledgeability. In order to do so, we will need to introduce a modification to the model of Section 3.3. In that version of the model, we assumed that the speaker had no uncertainty about the true world, i.e. that the speaker was fully knowledgeable. We will thus need to introduce the possibility of speaker ignorance into the model, and show how to integrate this with the rest of our machinery.

3.6.1 Modeling speaker ignorance

Our model of speaker ignorance builds on that of [36]. We assume that the speaker has made an observation o, which provides some degree of information about the world they are in. There is a joint prior distribution $P(w,o)$ over pairs of worlds w and observations o; when the speaker makes an observation, this induces a posterior distribution over worlds $P(\cdot|o)$.

When the listener hears an utterance u_p, they perform joint inference over the world w and the speaker's observation o. The definition of $K(w|u_p,s)$ remains the same as in

Equation 3.4, while the literal listener is now defined by:

$$L_0(w,o|u_p,s) \propto P(w,o)K(w|u_p,s) \tag{3.12}$$

This differs from the previous definition of the literal listener, in Equation 3.5, only in that the listener is now assigning a probability to a world-observation pair, and weighs this pair by the joint prior probability $P(w,o)$.

Our definition of the speaker differs most from Section 3.3. The speaker now potentially has uncertainty about the true state of the world, in addition to a QUD which imposes an equivalence relation on worlds. In order to properly define the speaker, we need to show how to combine these two features. As before, a QUD q is a function which maps a world w to its cell in a partition on worlds. Given a QUD q (which will often be left implicit), we define the marginal probability of a partition cell by:

$$P(q(w),o) = \sum_{w'} \mathbb{1}_{q(w')=q(w)} P(w',o) \tag{3.13}$$

This is the sum of probability of all worlds in the same partition cell as w. We next define the following equivalence relation on observations:

$$o \equiv_q o' \iff \forall w \; P(q(w)|o) = P(q(w)|o') \tag{3.14}$$

Under this definition, observations o and o' are equivalent when they induce the same distributions over partition cells. The QUD defines an equivalence relation on worlds, and this definition lifts this equivalence relation to observations.

We now define:

$$R_n(u_p,s|q(w),o,q) = \sum_{w',o'} \mathbb{1}_{(q(w')=q(w)) \wedge (o \equiv_q o')} L_n(w',o'|u_p,s) \tag{3.15}$$

This is the joint probability that the listener assigns to partition cell $q(w)$ and observations which are equivalent to o (under the equivalence relation $\equiv_q$). The speaker wants their utterance to be informative about the answer to the QUD, i.e. they want the listener to assign high probability to the correct cell of the partition. However, the speaker is potentially uncertain about which partition cell is the true one. In order to compute the informativeness of an utterance u_p, they therefore compute the average amount of information that the utterance provides about each QUD cells:

$$I_n(u_p, s|o,q) = -\sum_{q(w)} P(q(w)|o) \log \frac{1}{R_n(u_p, s|q(w), o, q)} \tag{3.16}$$

We use the notation $\sum_{q(w)}$ to indicate that the sum ranges over cells of the partition induced by QUD q.

The remaining speaker definitions are straightforward generalizations of those in Section 3.3:

$$\mathbb{E}_{P_N(\cdot|u_i,s)} I_n(\cdot|o,q) = \sum_{u_p} P_N(u_p|u_i,s) I_n(u_i, s|o,q) \tag{3.17}$$

$$U_n(u_i, s|o,q) = \mathbb{E}_{P_N(\cdot|u_i,s)} I_{n-1}(\cdot|o,q) - c(u_i) - c(s) \tag{3.18}$$

$$S_n(u_i, s|o,q) \propto e^{\lambda U_n(u_i,s|o,q)} \tag{3.19}$$

The pragmatic listener is only minimally modified from the previous definition. This listener has a model of how the speaker chooses utterances given an observation o and QUD q, and a prior distribution $P(w,o)$ over world-observation pairs. The listener performs joint inference over the world, the speaker's observation, and the QUD:

$$L_n(w, o, q|u_p, s) \propto P(w,o) P(q|o) \sum_{u_i} S_n(u_i, s|o,q) P_N(u_p|u_i, s) \tag{3.20}$$

3.6.2 Deriving knowledgeability inferences

This model can be used to derive the inference that stress indicates greater speaker knowledgeability. We will explain how this inference is derived in Example (20). For this example, we assume that there are two worlds, $\forall$ and $\exists\neg\forall$, corresponding to worlds in which all of the students passed the test and in which some but not all passed, respectively. We assume that there are three possible observations: the speaker could have observed that they are in world $\forall$, observed world $\exists\neg\forall$, or not gained any information about the world (in which case they have a uniform distribution over worlds). These observations correspond to the knowledge states $\{\forall\}$, $\{\exists\neg\forall\}$, and $\{\forall, \exists\neg\forall\}$, respectively, where each knowledge state is the set of worlds that the speaker considers possible. For example, in knowledge state $\{\forall, \exists\neg\forall\}$, the speaker is ignorant, and considers both worlds possible. The listener has a uniform prior distribution over the speaker's observation. The speaker wants to communicate exactly which world is the true one, and there is no uncertainty about this QUD.

There are two alternative utterances, *some* and *all*, which are given their normal semantics. Both utterances are assigned cost 1. The noise distribution over utterances is illustrated in Figure 3-6. The baseline probability of noise p is set to 0.01. The speaker can place stress on their utterance, which reduces the noise rate by a factor of 2. The use of stress is assigned cost 0.1.

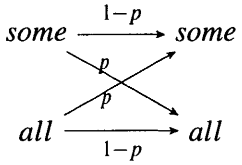


Figure 3-6: The noisy channel for Example (20), when stress is not used.

Figure 3-7 shows the predicted interpretation of stress in this example. When stress is placed on utterance *some*, the listener is more likely to draw the strong implicature: that the speaker knows that they are in world $\exists\neg\forall$. When stress is not placed on this utterance, the listener draws the weaker implicature: that the speaker does not know that world $\forall$ is true, and that they are in the ignorant knowledge state $\{\forall, \exists\neg\forall\}$.

To explain this, we first consider the speaker who is in knowledge state $\{\exists\neg\forall\}$, i.e. who knows that not all of the students passed the test. This speaker will almost always choose utterance *some*, as the only alternative, *all*, is literally incompatible with their intended meaning. In deciding whether to use stress on their utterance, this speaker will consider the consequences of having their intended utterance misheard by the listener. If the listener mishears *some* as *all*, then the listener will believe that they are in world $\forall$ with high probability, something which the speaker knows to be false. The speaker has a strong incentive not to communicate things which they know to be false, and will want to reduce the noise

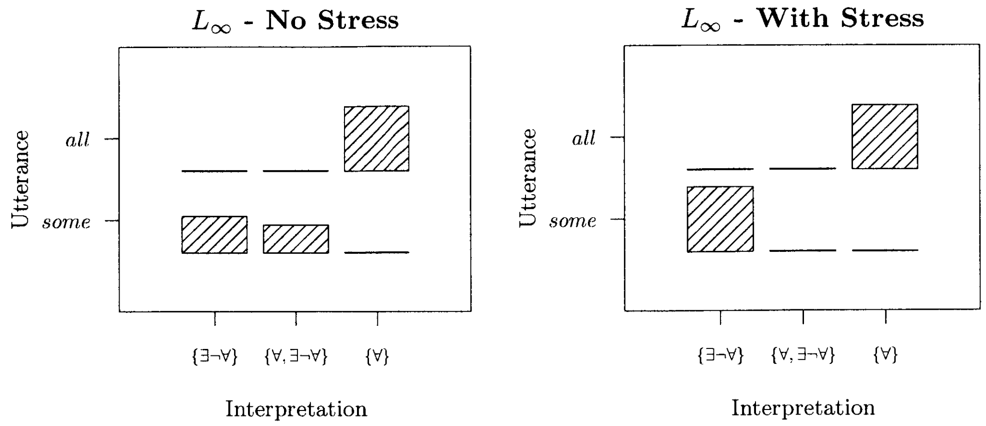


Figure 3-7: The listener's interpretation of the utterances without stress (on the left panel) and with stress.

rate on their utterance. This speaker will therefore be likely to place stress on *some*.

Next consider the speaker who is in the ignorant knowledge state $\{\forall, \exists\neg\forall\}$. Like the first speaker, this one will also choose utterance *some* with high probability. The speaker choice rule for this model, as defined in Equations 3.16-3.19, encodes a strong preference for the speaker to only choose utterances which they know to be true.[6] The speaker therefore needs to determine whether to place stress on *some*. Suppose that the listener mishears *some* as *all*. In this case, the listener will believe that world $\forall$ is true. The speaker does not know that this is the true world, but neither does this speaker know that it is not. This is distinct from the case of the speaker who knows that the true world is $\exists\neg\forall$, considered above. In that case, if the listener accidentally hears *all*, then they will have communicated something which they know to be false. There is an asymmetry between the two speaker types, corresponding to the difference between communicating something known to be false, and communicating something not known to be true (but still considered possible). It is possible to show that this asymmetry is encoded in the model, and has the consequence that noise decreases utility more for the knowledgeable speaker than for the

[6] More precisely, it encodes a preference for the speaker to only choose utterances which communicate answers to the QUD which they know to be correct. These constraints are equivalent in the current example, because the speaker's QUD is assumed to be maximally fine-grained.

ignorant speaker. As a result, the ignorant speaker has less incentive to reduce the noise rate of their utterance, and will be less likely to place stress on *some*.

The listener, in turn, knows that the speaker is more likely to place stress on *some* when they are knowledgeable than when they are ignorant. When the listener hears stress on *some*, they will infer that the speaker is more likely to be knowledgeable, i.e. they will be more likely to draw a strong implicature from *some*. This reasoning is reinforced at higher recursion-depths: the speaker knows that the listener is likely to interpret stress as indicating knowledgeability, and will therefore be more likely to use stress to communicate knowledgeability; the listener knows that the speaker knows this, and so on.

3.7 Disagreement and surprisal

We have so far explained several qualitative features of the interpretation of prosodic stress. First, we have explained why stress indicates that the speaker has a list-QUD rather than a polar-QUD. Second, we have explained why stress location can be used to indicate which specific list-QUD the speaker is trying to answer. We have further explained why the association between stress and QUD results in the exhaustive interpretation of stressed utterances: the listener infers that the speaker is providing a complete answer to the question that they are trying to answer.

There are certain features of stress interpretation which do not appear to be mediated by inferences about the QUD. In particular, stress can be used to indicate which parts of the utterance the speaker believes will be considered most surprising by the listener. Consider the following dialogue:

(21) A: Alice went to the store.
B: BOB went to the store.

Speaker B places stress on “Bob" in order to indicate disagreement with Speaker A's utterance. The speaker communicates that the locus of disagreement is who went to the store: the speaker agrees that someone went to the store, but it was not Alice. Contrast this with the next dialogue:

(22) A: Alice went to the store.

B: Alice went to the RESTAURANT.

Here, by placing stress on "Restaurant," the speaker communicates that the locus of disagreement is where Alice went: Alice went somewhere, but it was not the store.

More generally, speakers will use stress to indicate the parts of an utterance which they believe their interlocutors will find surprising.

(23) Would you believe it? BOB went to the store.

(24) Would you believe it? Bob went to the STORE.

In Example (23), the placement of stress on "Bob" indicates that it is surprising that Bob, rather than someone else, went to the store. That is, it indicates that the speaker believes that the listener has a prior expectation that a) Bob did not go to the store; b) someone else besides Bob would have been more likely to have gone to the store. It does not indicate that the listener has a prior expectation that Bob would have been likely to have gone somewhere else besides the store. In Example (24), the placement of stress on "store" indicates that it is surprising that Bob went to the store, rather than somewhere else. That is, the speaker believes that the listener has a prior expectation that a) Bob did not go to the store; b) it would have been more likely for Bob to have gone somewhere else instead of the store. It does not indicate that the listener has a prior expectation that someone else besides Bob would have been more likely to have gone to the store.

The noisy-channel pragmatics model predicts these effects. These effects are not mediated by listener inferences about the QUD; in fact, we will assume a deterministic (trivial) QUD throughout our discussion in this section. These phenomena will instead be derived through a distinct reasoning process.

3.7.1 Stress and disagreement

We will first present a model of scenarios such as Example (21). In this example, the speaker uses stress to signal disagreement with a prior assertion in the discourse. Our modeling will in fact apply to a more general class of scenarios, in which it is common knowledge that some proposition is assigned low prior probability by the listener. This includes cases in which the listener has made a prior assertion in the discourse which contradicts this proposition, but it also includes cases in which the proposition is known to have low prior probability for some other reason, e.g. due to its obvious *a priori* implausibility. This includes examples such as:

(25) A: I heard that the theorem was proved recently.
B:The LITTLE BOY proved it.

In such cases, the speaker uses stress to indicate that they are saying something which they know will be considered *a priori* implausible.

Disagreement without compositional structure

Our modeling setup will be similar to the examples in Sections 3.4 and 3.5, with two primary differences. First, we do not assume a uniform prior distribution over worlds. Instead, there will be a distinguished world which is assigned high prior probability by the listener; the speaker will know that this world does not hold, and will use stress to signal disagreement with the listener's prior expectations. Second, we assume that the listener has no uncertainty about the speaker's QUD. The speaker is assumed to always want to answer the question, *What happened?* Given this QUD, the speaker wants to communicate to the listener the exact world that they are in.[7] Nothing in our model derivations depend on this strong of an assumption, but more complex QUDs are not essential for understanding the example.

For simplicity, we assume that there are two worlds, $\{Alice\}$, in which only Alice went to the store, and $\{Bob\}$, in which only Bob went to the store. The speaker knows who went to the store, and listener assigns prior probability $P(\{Alice\}) > 0.5$ to world $\{Alice\}$.

There are two alternative utterances, a and b, which receive the same semantics as in Section 3.4. Each utterance is assigned cost 1. Noise is assumed to occur with probability

[7] Note that this QUD is implicitly assumed in prior work using the RSA model [28].

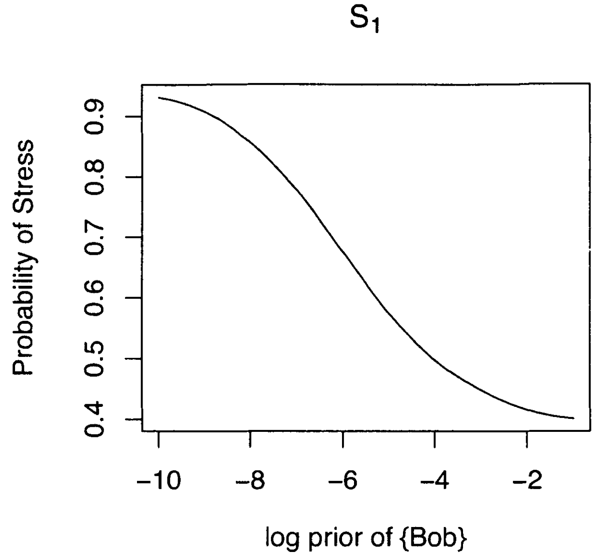


Figure 3-8: The probability that the speaker who wants to communicate world $\{Bob\}$ will use stress, as a function of the prior probability of this world.

p, and the noise distribution is the same as in Figure 3-1. Prosodic stress has the same effect as in the previous examples: it reduces the probability of noise by a factor of 2. Stress is assigned cost 0.1.

Suppose that the speaker wants to communicate world $\{Bob\}$, and chooses utterance b to do so. Figure 3-8 shows the probability that the speaker will use prosodic stress as a function of the prior probability $P(\{Alice\})$ that the listener assigns to world $\{Alice\}$. Higher values of $P(\{Alice\})$ result in higher probability of the speaker choosing to use prosodic stress. That is, if the listener has lower belief in the correct world, the speaker is more likely to place prosodic stress on their utterance.

To understand why this occurs, consider the listener who assigns high prior probability to world $\{Alice\}$, and who hears utterance b without prosodic stress. This utterance literally communicates that Bob went to the store, something which the listener believes to be unlikely. There are two options for this listener: a) the utterance was actually intended by the speaker, and the speaker wanted to communicate something *a priori* unlikely; or b) the utterance was corrupted by noise, and the speaker actually wanted to communicate the more likely world $\{Alice\}$. When world $\{Alice\}$ is sufficiently probable *a priori*, then the possibility of noise becomes more likely than the possibility that the speaker actually intended world $\{Bob\}$, and the rational inference for the listener is b). If, however, the speaker has placed stress on their utterance, then the situation is somewhat different. The listener knows that a stressed utterance is less likely to have been corrupted. As a result, the listener will be less likely to infer option b). The listener is instead more likely to infer that the speaker really intended to communicate something *a priori* unlikely. For sufficiently high values of the prior probability $P(\{Alice\})$, the listener will still infer that the utterance was corrupted. However, prior probability threshold is higher when the speaker has used stress than when they haven't.

Next consider the speaker who wants to communicate world $\{Bob\}$. This speaker will clearly choose utterance *b*, as this is the only utterance that provides information about their intended world. The speaker knows the listener's prior probability assignments, and how the listener will interpret their utterance given this prior distribution. For sufficiently high prior probability assigned to $\{Alice\}$, the speaker knows that their unstressed utterance will be interpreted as world $\{Alice\}$. The listener will assign lower probability to this world if they use stress. As a result, the speaker in this case will use stress to ensure that their utterance is interpreted correctly.

Introducing compositional structure

We have shown how the model derives the use of stress to signal disagreement in simple cases such as Example (21). We have so far only discussed an idealized setting in which the choice to use stress is a binary decision, and stress signals global disagreement with prior discourse assumptions. However, the contrast between Examples (21) and (22), reproduced and expanded below, demonstrates that the placement of stress within a sentence can convey information about what is being disagreed with:

(26) A: Alice went to the store.
B: a. No, BOB went to the store.
b. # No, Bob went to the STORE.

(27) A: Alice went to the store.

B: a. No, Alice went to the RESTAURANT.

b. # No, ALICE went to the restaurant.

If the speaker wants to disagree with the claim that Alice, rather than Bob, went to the store, then they will signal this by placing stress on "Bob" rather than "store." In contrast, if the speaker wants to disagree with the claim that Alice went to the store, rather than the restaurant, they will stress "restaurant" rather than "Alice." Our model must therefore be able to explain why the choice of stress location provides information about which part of the prior discourse is being disagreed with.

We will extend the previous example in order to model these inferences. We assume that there are two individuals, Alice and Bob, and two locations, the store and the restaurant. In every world, one individual went to one location, and nobody went anywhere else. There are four worlds which fit these constraints. Using the notation of Section 3.5.1, these worlds are denoted by $\{\{Alice\}_S\}$ (in which Alice only went to the store, and Bob did not go anywhere), $\{\{Alice\}_R\}$, $\{\{Bob\}_S\}$, and $\{\{Bob\}_R\}$. We assume that the prior probability of world $\{\{Alice\}_S\}$ is greater than that of every other world, and that the other worlds are assigned equal probability. This represents the listener's belief that Alice went to the store. The worlds $\{\{Alice\}_R\}$ and $\{\{Bob\}_S\}$ will be the crucial ones for this example. They each differ from the high-probability world $\{\{Alice\}_S\}$ in a single dimension. In $\{\{Alice\}_R\}$ and $\{\{Alice\}_S\}$, Alice went somewhere; the worlds are distinguished only with respect to where Alice went (the restaurant or the store). In $\{\{Bob\}_S\}$ and $\{\{Alice\}_S\}$, someone went to the store; these worlds are distinguished only with respect to who went to the store (Bob or Alice).

There are four alternative utterances: $a : s$ ("Alice went to the store"), $a : r$, $b : s$, and $b : r$. These utterances are assigned the same semantics as in Section 3.5.1. The speaker has the option to place stress on the left phrase in the utterance, the right phrase, or on neither of the phrases. Utterance cost, prosodic stress cost, and the effect of prosody on the noise rate remain the same as in that section. The noise distribution remains the same as well, and is shown in Figure 3-3.

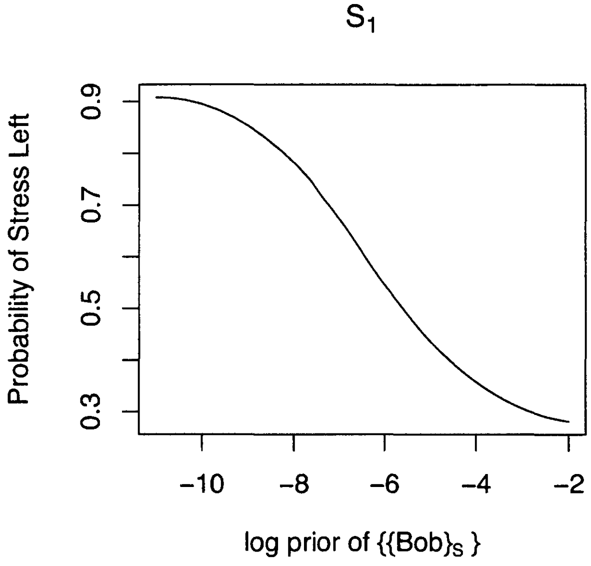


Figure 3-9: The probability that the speaker who wants to communicate world $\{\{Bob\}_S\}$ will use stress $Left$, as a function of the prior probability of this world.

Figure 3-9 shows the model's predictions about the use of prosodic stress, as a function of prior probability. The figure shows the predictions for the speaker who wants to communicate world $\{\{Bob\}_S\}$ (corresponding to Example (26)); the predictions are symmetric for the speaker who wants to communicate world $\{\{Alice\}_R\}$ (corresponding to Example (27)). As $P(\{\{Alice\}_S\})$ increases, and $P(\{\{Bob\}_S\})$ correspondingly decreases, it becomes more likely that the speaker in world $\{\{Bob\}_S\}$ will place stress on the left phrase of their utterance ("Bob"). By symmetry, it becomes more likely that the speaker in world $\{\{Alice\}_R\}$ will place stress on the right phrase ("restaurant").

We will explain why the speaker who wants to communicate world $\{\{Bob\}_S\}$ will place stress on the left phrase of their utterance, "Bob." A symmetrical argument will explain why the speaker in world $\{\{Alice\}_R\}$ will place stress on the right phrase. The speaker wants to communicate that Bob went to the store, and will clearly choose the utterance $b : s$, as this is the only utterance literally compatible with this world. In deciding where in the utterance to place stress, this speaker will reason about how the listener who hears utterance $b : s$ will interpret this utterance.

By assumption, the listener assigns high prior probability to world $\{\{Alice\}_S\}$. When the listener hears $b : s$, they will have two options: 1) they can infer that $b : s$ is really the intended utterance of the speaker, and that the speaker wanted to communicate the low-probability world $\{\{Bob\}_S\}$; or 2) they can infer that the speaker intended some other utterance, and wanted to communicate the higher probability world $\{\{Alice\}_S\}$. If the listener infers 2), then their inference is that the speaker actually intended utterance $a : s$ (as this is the only utterance literally compatible with world $\{\{Alice\}_S\}$), and that the phrase a was corrupted to b. For sufficiently high prior probability of world $\{\{Alice\}_S\}$, this is the rational inference for the listener to make. If the speaker has placed stress on the right phrase s, then this does not change the listener's inferences. Placing stress on s decreases the probability that the listener mistakenly heard this phrase. However, the listener already believes that they have accurately perceived the phrase s; in both scenarios 1) and 2), the listener believes that the speaker intended phrase s, and their only source of uncertainty is regarding the first phrase in the speaker's utterance. Thus, reducing the noise rate on s does not provide evidence for 1) over 2), or vice-versa. In contrast, placing stress on the left phrase b does change the listener's inferences. If the speaker believes that the utterance was more likely to have been transmitted accurately, then they will be more likely to infer that the perceived utterance $b : s$ was actually intended, and that the speaker wanted to communicate the low-probability world $\{\{Bob\}_S\}$.

The speaker knows that placing stress on b will make the listener more likely to correctly interpret the utterance, while placing stress on s will not. As a result, this speaker will be more likely to place stress on b.

3.7.2 Stress and surprisal

We have so far explained why the model predicts that stress can be used to indicate disagreement with the listener's *a priori* expectations, when these expectations are common knowledge among the speaker and listener. We will now address a more complex type of scenario, in which the speaker uses stress to communicate both that some event occurred, *and that it is surprising*. This is illustrated by Examples (23) and (24), which are reproduced below:

(28) Would you believe it? BOB went to the store.

(29) Would you believe it? Bob went to the STORE.

In Example (28), the speaker uses stress to indicate that it is surprising that Bob, rather than someone else, went to the store. The speaker can felicitously use this utterance even if the listener does not know that it was unlikely for Bob to have gone to the store (and more likely for someone else to have gone). In such a case, the speaker is communicating both that Bob went to the store, and that the event was *a priori* unlikely. The second type of information communicated can be considered a type of accommodation: the listener gains information about the speaker's beliefs about the discourse context.

In order to model this type of accommodation, we will need to modify the model introduced in Section 3.3. In that version of the model, it was assumed that the listener's prior distribution over worlds was common knowledge among the speaker and listener. In the current version of the model, we will assume that the speaker has beliefs about the listener's prior distribution over worlds, but that these beliefs are potentially unknown by the listener. Thus, in order to interpret the speaker's utterance, the listener will need to perform joint inference over both what world the speaker believes they are in, and what the speaker believes the listener's prior expectations are.

The definition of the literal listener remains largely the same as in Section 3.3, except that the listener is now parameterized by a prior distribution P_k:

$$L_0(w|u_p,s,P_k) \propto P_k(w)K(w|u_p,s) \tag{3.21}$$

This prior distribution may potentially vary from listener to listener. The definition of $K(w|u_p,s)$ remains unchanged from Equation 3.4.

The speaker is also defined in largely the same way. The only change is that the speaker is now parameterized by the prior distribution P_k, which determines the speaker's beliefs about the listener's prior distribution:

$$R_n(u_p, s|w, q, P_k) = \sum_{w'} \mathbb{1}_{q(w')=q(w)} L_n(w'|u_p, s, P_k) \tag{3.22}$$

$$I_n(u_p, s|w, q, P_k) = -\log \frac{1}{R_n(u_p, s|w, q, P_k)} \tag{3.23}$$

$$\mathbb{E}_{P_N(\cdot|u_i,s)} I_n(\cdot|w, q, P_k) = \sum_{u_p} P_N(u_p|u_i, s) I_n(u_i, s|w, q, P_k) \tag{3.24}$$

$$U_n(u_i, s|w, q, P_k) = \mathbb{E}_{P_N(\cdot|u_i,s)} I_{n-1}(\cdot|w, q, P_k) - c(u_i) - c(s) \tag{3.25}$$

$$S_n(u_i, s|w, q, P_k) \propto e^{\lambda U_n(u_i, s|w, q, P_k)} \tag{3.26}$$

The most substantial change to the model is in the definition of the pragmatic listener L_n. This listener has a prior distribution over worlds P_j. However, this prior distribution may not coincide with the distribution P_k which the speaker believes that the listener is using. The listener knows this, and is uncertain about which prior distribution the speaker believes that the listener is using:

$$L_n(w, q, P_k|u_p, s, P_j) \propto P_j(w) P(q|w) P(P_k) \sum_{u_i} S_n(u_i, s|w, q, P_k) P_N(u_p|u_i, s) \tag{3.27}$$

The term $P(P_k)$ is the prior probability that the listener assigns to prior distribution P_k. The marginal probability of world w and QUD q can be computed by:

$$L_n(w, q|u_p, s, P_j) \propto P_j(w) P(q|w) \sum_{k} P(P_k) \sum_{u_i} S_n(u_i, s|w, q, P_k) P_N(u_p|u_i, s) \tag{3.28}$$

Given these changes to the model, we can now explain why the use of stress in examples such as (28) can be used to communicate that something surprising occurred. We will consider the simplest possible example of this reasoning; more complex cases will follow from similar reasoning.

We will assume that there are two worlds, as in Section (25): $\{Alice\}$ and $\{Bob\}$, corresponding to who went to the store. There are three possible prior distributions over these worlds:

$$P_1(\{Alice\}) = 0.5,\ P_1(\{Bob\}) = 0.5$$
$$P_2(\{Alice\}) = p,\ P_2(\{Bob\}) = 1 - p$$
$$P_3(\{Alice\}) = 1 - p,\ P_3(\{Bob\}) = p$$

We assume that $p > 0.5$, so that P_2 assigns higher probability to $\{Alice\}$, while P_3 assigns higher probability to $\{Bob\}$. The distribution P_1 assigns equal probability to both worlds. We assume that the speaker always knows which world is the true one. Each speaker is parameterized by one of the prior distributions. This represents the speaker's beliefs about the listener's prior expectations. For example, the speaker parameterized by P_2 believes that the listener expects $\{Alice\}$ to be more likely than $\{Bob\}$. The pragmatic listener, in turn, is uncertain about which prior distribution the speaker is parameterized by. We assume that the listener has a uniform prior over the speaker's prior distribution parameter.

The rest of our modeling setup is the same as in Section (25). There are two utterances, a and b, with our standard cost and noise assumptions. The speaker is assumed to have a maximally precise QUD, i.e. they want to communicate exactly which world they are in.

When the speaker places stress on utterance a, the listener is more likely to infer that the speaker is parameterized by prior distribution P_3, the distribution which assigns lower probability to world $\{Alice\}$. That is, when the speaker chooses utterance a with stress, the listener infers two things: that Alice went to the store (rather than Bob), and that the speaker believes that this is surprising. By symmetry, when the speaker chooses utterance b with stress, the listener infers that the speaker believes that it is surprising that Bob went to the store. In addition, as the distribution P_3 assigns less probability to world $\{Alice\}$, the listener becomes more confident that the use of stress on utterance a signals this distribution.

The model makes these predictions for reasons which are closely related to its predictions in Section (25). In that section, we explained why the model predicts that stress signals disagreement, i.e. why the speaker will use stress when they are conveying information which they know the listener assigns low prior probability. The same reasoning applies in the current scenario. Suppose that the speaker wants to communicate world $\{Alice\}$, and is parameterized by prior distribution P_3. This speaker believes that the listener assigns low prior probability to the true world. As a result, as shown in Section (25), this speaker will choose utterance a and will place stress on this utterance. They know that they are communicating something surprising to the listener, and want to ensure that the listener does not mistakenly infer that they were trying to communicate something more *a priori* likely. In contrast, the speaker who wants to communicate world $\{Alice\}$, and who is parameterized by prior distribution P_2, will be unlikely to place stress on their utterance. This speaker believes that the listener assigns high prior probability to the correct world. Because the utterance a will only further confirm the listener's prior expectations, this speaker believes that the utterance will be interpreted correctly even without the use of stress.

Next, consider the inferences of the listener who hears utterance a with stress. We will assume that this listener has prior distribution P_1, i.e. that this listener assigns equal probability to the two worlds. This listener knows that the speaker would have been likely to use stress if they had been parameterized by distribution P_3, but unlikely to use it if they were parameterized by distribution P_2. As a result, this listener will infer that the speaker is more likely to have been parameterized by P_3. That is, the listener will infer that the speaker believed that they were communicating something surprising to the listener.

3.8 Strengthening of underspecified utterances

In this section, we consider several interactions between stress and the interpretation of semantically underspecified lexical items. The effects of stress that we consider in this section are qualitatively distinct from those considered in the previous sections. In particular, these effects are not driven by inferences about the QUD, as in Section 3.4, or by inferences about prior information in the discourse context, as in Section 3.7. These effects have additional significance, as they cannot be straightforwardly derived within standard approaches to stress interpretation, such as Alternative Semantics [84].

We will first consider semantic underspecification in the context of gradable adjectives, such as *tall*, *big*, and *expensive*. These adjectives are each associated with a scale, for example a height scale in the case of *tall* [54, 55]. When applied to an argument, the adjectives conveys that the argument falls above (or below) some threshold on this scale. For example, *tall* conveys that its argument is above some height threshold:

(30) Bob is tall. [Understood meaning: The man's height is greater than normal, e.g. it is above 74 inches.]

An important property of such adjectives is that their interpretation is context-dependent. In particular, an adjective's scale threshold is generally not fixed prior to the discourse; rather, it is sensitive to the distributional properties of the relevant contrast-class, and to the interests of the participants in the conversation [57, 55, 3]. Consider a situation in which Bob is a professional basketball player (and the speaker and listener have been discussing other basketball players):

(31) Bob is tall. [Understood meaning: Bob is taller than normal basketball players, e.g. his height is greater than 82 inches.]

When *tall* is used in the context of adult men, it conveys that the individual is tall relative to other adult men. When it is used in the context of professional basketball players, it conveys that the individual is tall relative to other basketball players. In both cases, the listener needs to infer the speaker's intended comparison class, and use the distributional properties of this comparison class to infer the intended threshold. Gradable adjectives such as *tall* are thus semantically underspecified: the scale (e.g. height) and structure (e.g. lower-boundedness) of each adjective's meaning is fixed in advance, but the precise value of the threshold is not.

Stress can be used to systematically strengthen the interpretation of gradable adjectives. When the adjective is used predicatively, stress will often signal that an adjective should receive an intensified interpretation, i.e. that its threshold should be especially high (or low, depending on the direction of the adjective):

(32) Bob is TALL.

(33) Alice is SMART.

(34) The watch is EXPENSIVE.

In these examples, we assume that the utterance is not being used contrastively, e.g. that someone in the discourse has not previously asserted that Bob is short. Given this assumption, the use of stress in Example (32) indicates that Bob is especially tall — taller than would be expected from Example (30) — and similarly in the other examples. As these examples illustrate, the effect appears to be quite general, occurring whenever stress applied to a lower- or upper-bounding adjective in a predicative setting.

In addition to gradable adjectives, we will consider the effects of stress on a different class of semantically underspecified lexical items, quantifiers with underspecified domains. We will be focusing in particular on universal quantifiers:

(35) Every girl came to the party.

(36) Bob always arrived at work on time.

In certain respects, the universal quantifiers *every* and *always* have a fixed, clearly specified semantics: each ranges over a domain (of individuals and events/time-points, respectively), and predicates each member of this domain with some property. However, the domain of quantification is not fixed in advance, as part of the semantics of these lexical items [59, 4, 98, 67]. The listener must infer the quantifier domain from the discourse context: from what is salient in the context, and what they believe the speaker wants to talk about. In Example (35), the speaker is not asserting that every girl in the world came to the party. Rather, the speaker is asserting that some contextually relevant set of girls — e.g., all of the girls in the third grade — came to the party. In Example (36), the speaker is not asserting that Bob arrived at work on time at every time point, or even that Bob arrived at work on time every single work day. Rather, they are asserting that, for some contextually relevant set of work days, Bob arrived at work on time on those days.

When stress is placed on one of these universal quantifiers, it signals an expansion of the quantifier domain:

(37) EVERY girl came to the party.

(38) NOBODY brought presents.

(39) The bus went NOWHERE.

(40) Bob ALWAYS arrived at work on time.

In Example (37), the use of stress indicates that *every* quantifies over a larger set of girls than in the unstressed Example (35). For example, consider a scenario in which some girls in the third grade class did not receive invitations to the party. The unstressed utterance may indicate that all of the invited girls came to the party, while the stressed utterance may indicate that the uninvited girls (surprisingly) came too. In Example (40), the use of stress indicates that *always* quantifies over more time points than in Example (36). The unstressed Example (36) may indicate that Bob routinely arrived at work on time, while the stressed Example (40) may indicate that Bob was unusually punctual, e.g. that he even showed up for work on time on holidays or following natural disasters.

Both gradable adjectives and quantifiers are semantically underspecified: gradable adjectives have a free parameter which specifies a scale threshold, while quantifiers have a free parameter which specifies the domain of quantification [68, 98]. For both types of lexical items, the application of stress strengthens the interpretation of the utterance, either by pushing the threshold parameter towards a more extreme value for the adjectives, or by expanding the size of the domain for universal quantifiers.

These effects have theoretical interest, as they do not have straightforward derivations within an Alternative Semantics approach to stress interpretation. Under such an account, stress is used to indicate the focus of an utterance. The placement of focus, in turn, determines the set of alternatives to the utterance. Any effect of stress placement on interpretation will be mediated by these alternatives. In particular, the effects of stress in simple declarative utterances (in the absence of any focus-sensitive operators) will typically be derived through exhaustification of the alternatives: the listener infers that none of the alternatives are true. In Example (32), the listener may consider the expression *Bob is short* as an alternative, and take the negation of this. But this will not derive a strengthened reading of the utterance: the literal meaning of Example (32) already communicates that Bob is not short. Similarly, for Example (37), exhaustifying with respect to the obvious alternatives will not derive a strengthened interpretation of the utterance. This utterance will clearly have *Most girls came to the party* and *Some girls came to the party* as alternatives. But taking the negation of these alternatives will lead to a contradiction: it is not possible for all of the girls to have gone to the party, if it is not the case that some (or most) did. Thus, it does not appear that exhaustification is sufficient for deriving the interpretation of stress in these cases.

3.8.1 Inference over free semantic variables

Here, we will explain how to integrate semantic underspecification into the noisy-channel pragmatics model. In order to specify this model, we first need to fix a representation of semantic underspecification. We will follow much previous work in using a *free variable* representation of underspecification [66, 72]. Under this approach, an underspecified expression is parameterized by a free variable. This variable supplies all of the information which is necessary for interpreting the expression. The lexical entry for the expression does not contain a fixed value for the variable; rather, this value needs to be filled in during pragmatic inference.

In general, an expression A with a free semantic variable is represented by [60]:

$$[\![A]\!] = \lambda\theta\lambda x[F_A(x,\theta)] \tag{3.29}$$

The variable θ is the free variable in this expression, and F_A is a semantic interpretation function for a. For example, the lexical item *tall* will be represented by:

$$[\![tall]\!] = \lambda\theta_{tall}\lambda x[\mu_{height}(x) \geq \theta_{tall}] \tag{3.30}$$

The term θ_{tall} is the free variable for *tall*, which takes on real-numbered values; an object is tall if its height is greater than the value of θ_{tall}. The lexical representation for the quantifier *all* is:

$$[\![all]\!] = \lambda\theta_{all}\lambda X\lambda Y[(X \cap \theta_{all}) \subset Y] \tag{3.31}$$

The free variable θ_{all} specifies the quantifier domain. For sets A and B, the expression $((([\![all]\!](\theta_{all}))(A))(B)$ is true iff all of the mutual elements of A and θ_{all} are also elements of B. We will use the notation $[\![A]\!]^{\theta}$ to denote the interpretation of A after the value of θ has been fixed:

$$[\![A]\!]^{\theta} = \lambda\theta\lambda x[F_A(x,\theta)](\theta) = \lambda x[F_A(x,\theta)] \tag{3.32}$$

If the expression A contains multiple semantic free variables, then we will use the vector θ to represent the value of these variables, and we will denote the interpretation of A with respect to these values by $[\![A]\!]^{\theta}$.[8]

We model the resolution of the semantic free variables as a statistical inference problem. In particular, we follow the reformulation of lexical uncertainty [] proposed in []. Under this model, the literal listener and the speaker S_1 have no uncertainty about the values of the semantic free variables. That is, they have fixed beliefs about the literal interpretation of each utterance, e.g. about which height threshold qualifies an object as being tall. All of the uncertainty is lifted to the listener L_1. This listener is *a priori* uncertain about values of the variables. They thus do not know how the speaker S_1 expects utterances to be literally interpreted. In order to infer the speaker's intended meaning, this listener must thus reason jointly about which meanings and semantic specifications would have made the speaker likely to have chosen the perceived utterance.

We will now present our formal model definitions. The literal listener is largely the same as in previous versions, with one exception. This listener is now parameterized by the vector θ, representing the values of all of the semantic free variables necessary for interpretation. We thus assume that the literal listener has no uncertainty about the value of the free variables:

$$K(w|u_p,s,\theta) = \sum_{u_i} P(u_i|u_p,s)\mathbb{1}_{w\in[\![u_i]\!]^{\theta}} \tag{3.33}$$

$$L_0(w|u_p,s,\theta) \propto P(w)K(w|u_p,s,\theta) \tag{3.34}$$

[8] Though we have not formally defined the interpretation of an expression with respect to multiple free variables, this can be done straightforwardly as an extension of our previous definitions.

In contrast to the previous versions of the model, we define the speaker S_1 and the higher-order speakers S_n ($n > 1$) separately. Like the literal listener, the speaker S_1 is parameterized by the vector θ, representing the values of the semantic free variables in the available alternative utterances. This speaker has no uncertainty about the values of these variables, and thus no uncertainty about how the literal listener will interpret their utterance. The speaker is defined by:

$$R_0(u_p,s|w,q,\theta) = \sum_{w'} \mathbb{1}_{q(w')=q(w)} L_0(w'|u_p,s,\theta) \tag{3.35}$$

$$I_0(u_p,s|w,q,\theta) = -\log \frac{1}{R_0(u_p,s|w,q,\theta)} \tag{3.36}$$

$$\mathbb{E}_{P_N(\cdot|u_i,s,\theta)} I_0(\cdot|w,q) = \sum_{u_p} P_N(u_p|u_i,s) I_0(u_i,s|w,q,\theta) \tag{3.37}$$

$$U_1(u_i,s|w,q,\theta) = \mathbb{E}_{P_N(\cdot|u_i,s)} I_0(\cdot|w,q,\theta) - c(u_i) - c(s) \tag{3.38}$$

$$S_1(u_i,s|w,q,\theta) \propto e^{\lambda U_1(u_i,s|w,q,\theta)} \tag{3.39}$$

The definition of listener L_1 represents the most substantial change from our previous models. This listener does not know the values of the semantic free variables in advance, but rather has a prior distribution $P(\theta)$ over these values. The joint probability of a world w, QUD q, and vector of semantic values θ is defined by:

$$L_1(w,q,\theta|u_p,s) \propto P(w)P(q|w)P(\theta) \sum_{u_i} S_1(u_i,s|w,q,\theta) P_N(u_p|u_i,s) \tag{3.40}$$

Using this joint distribution, the listener can compute the marginal probability of a world-QUD pair w and q, or of a vector of semantic values θ:

$$L_1(w,q|u_p,s) \propto P(w)P(q|w) \sum_{\theta} P(\theta) \sum_{u_i} S_1(u_i,s|w,q,\theta) P_N(u_p|u_i,s) \tag{3.41}$$

$$L_1(\theta|u_p,s) \propto P(\theta) \sum_{w} P(w) \sum_{q} P(q|w) \sum_{u_i} S_1(u_i,s|w,q,\theta) P_N(u_p|u_i,s) \tag{3.42}$$

The higher-order speakers S_n and listeners L_n (for $n > 1$) are defined exactly as in Section 3.3. Though inference over the value of the semantic variables is limited to the listener L_1, its effects are not limited to this listener; as we explain below, asymmetries arising from this listener are propagated through the inferences of the higher-order speakers and listeners.

3.8.2 Adjective strengthening

We will now explain why the use of stress is predicted to intensify the interpretation of gradable adjectives. We will specifically consider the use of stress in Example (32). We will assume that the speaker knows Bob's exact height, and that the speaker's QUD is *What is Bob's exact height?* The set of possible worlds is $\{1,...,n\}$, where each integer represents a possible height for Bob. The elements of this set are given their natural ordering, so that higher elements represent greater heights. For the current example, we set $n = 7$, but the qualitative predictions of the model are not sensitive to this choice. We assume that the speaker wants to communicate the exact value of Bob's height. The listener's prior distribution is uniform over worlds.

There are three alternative utterances: *tall*, *short*, and u_{null}. The semantics for *tall* is given in Equation 3.30, and the semantics for *short* is defined similarly:

$$[\![short]\!] = \lambda\theta_{short}\lambda x[\mu_{height}(x) \leq \theta_{short}] \tag{3.43}$$

The utterances *tall* and *short* are assigned cost 1. The utterance u_{null} is included in order to make the model well-defined; the technical issues motivating its inclusion are discussed further in [9]. This utterance is given a trivial semantics, so that it is true in every world, and is assigned very high cost, so that it is rarely used by the speaker and has minimal effect on pragmatic reasoning. Here it is assigned cost 10.

The two alternative utterances *tall* and *short* have semantic free variables θ_{tall} and θ_{short}, respectively, which specify their height thresholds. The listener L_1 must have a prior distribution over the values of these variables. We assume that each variable is independently and uniformly distributed across the possible heights in $\{1,...,n\}$. Thus the listener does not have any prior information about the height thresholds associated with the utterances.

The noise distribution is illustrated in Figure 3-10. The figure illustrates that the utterances *tall* and *short* may be confused for each other with probability p. As in previous examples, we set $p = 0.01$. We assume that the speaker may place prosodic stress on any of the utterances, decreasing the probability of noise by a factor of 2. Stress is assigned cost 0.1.

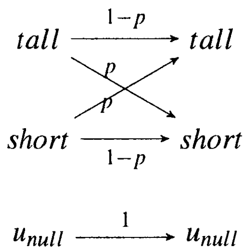


Figure 3-10: The noisy channel for Example (32), when prosodic stress is not used.

Figure 3-11 shows the effect of stress on utterance interpretation in the model. When the speaker places stress on *tall*, the listener infers that Bob has a greater height; when the speaker places stress on *small*, the listener infers that Bob has a smaller height.

In order to explain these effects, consider a speaker S_1 who believes that Bob has height 7, i.e. the maximal height in this example. This speaker will almost always choose utterance *tall*: this utterance almost always increases the probability of world 7, while the alternative *short* is almost always incompatible with this utterance.[9] Thus the speaker only needs to decide whether to place stress on *tall*. If the listener mishears *tall* as *short*, then they will almost always form beliefs which are incompatible with the speaker's intended meaning. For all values of $\theta_{short} < 7$, if the listener hears *short*, then they will conclude that Bob has height less than 7; this occurs with probability $\frac{6}{7}$. As a result, the speaker will have a strong incentive to transmit the utterance *tall* accurately, and will be relatively likely to use stress to reduce the noise rate.

Next consider a speaker S_1 who believes that Bob has a somewhat lower height, e.g. 5. This speaker will typically choose utterance *tall* rather than *short*. Whenever $\theta_{tall} \leq 5$, the utterance *tall* will be compatible with world 5. This occurs with probability $\frac{5}{7}$. In

[9] The only exception to the first claim is when $\theta_{tall} = 1$, and the only exception to the second is when $\theta_{short} = 7$.

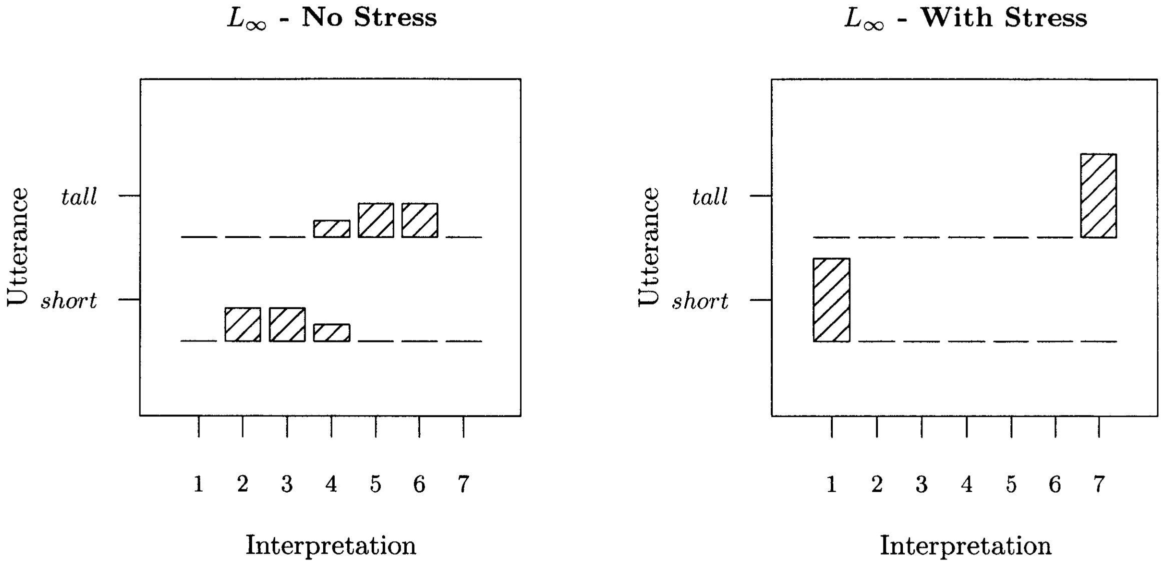


Figure 3-11: The listener's interpretation of the adjectives without stress (left panel) and with stress. The figures show the probability that the listener assigns to each height, as indicated on the x-axis.

contrast, the utterance *short* will be compatible with the world only when $\theta_{short} \geq 5$. This occurs with probability $\frac{3}{7}$. Thus, the utterance *tall* has a higher probability of being literally compatible with the world than the utterance *short*. Moreover, it is straightforward to show that, on average, *tall* will be more informative about this world than *short*. This means that, as in the previous case, the speaker will choose *tall* with high probability. How likely is this speaker to place stress on the utterance? If the listener mishears *tall* as *short*, then they will infer that Bob's height is less than or equal to θ_{short}. Thus, if $\theta_{short} < 5$, then they will infer that Bob's height is less than 5, and their beliefs will be incompatible with the speaker's intended meaning; this occurs with probability $\frac{4}{7}$. Thus, though the listener has a relatively high chance of forming beliefs which are incompatible with the intended world if they mishear the speaker's utterance, this occurs with lower probability than in the case that the speaker wants to communicate world 7. The speaker has less incentive to reduce the noise rate in this case than in the case that they want to communicate world 7, and therefore will be less likely to use stress.

We have so far explained why the speaker is more likely to place stress on *tall* if they are in world 7 (or, more generally, worlds in which Bob is especially tall) than if they are in world 5 (or, more generally, worlds in which Bob is less tall). We next consider the inferences of the listener L_1 who hears stress on *tall.* This listener knows that the speaker S_1 is more likely to place stress on this utterance if Bob is especially tall. As a result, if the utterance *tall* is stressed, the listener will infer that it is likely that the speaker wants to communicate that Bob is especially tall; if *tall* is unstressed, they will be less likely to believe this. This asymmetry between the stressed and unstressed utterance is strengthened by the higher-depth speakers and listeners. The speaker S_n (for $n > 1$) knows that the listener will interpret stress on *tall* as indicating a high value for Bob's height, and as a result is more likely to place stress on *tall* in order communicate greater heights. The listener L_n (for $n > 1$) knows that the speaker is more likely to use stress for greater values of Bob's height, and as a result will infer that stress indicates a greater height.

3.8.3 Quantifier strenghtening

In the previous section, we explained why the use of stress will strengthen the interpretation of gradable adjectives. In the current section, we explain why stress will strengthen the interpretation of certain quantifiers. We will first explain the model's predictions for Example (37), in which the use of stress expands the domain of the universal quantifier. We will then consider the model's predictions for other quantifiers, including *most* and *some.*

Modeling assumptions

We will assume that there was a party which was attended by some set of girls, and that worlds are specified by the set of girls who went to the party. The set of girls G is equal to $\{1,...n\}$, and therefore each world is specified by a subset of G. For the current example, we set $n = 6$, though the qualitative predictions of the model are robust to alternate values of n. We assume that at least one girl went to the party, so that the set of worlds $\mathcal{W} = 2^G \setminus \{\{\}\}$. The speaker knows who went to the party, and their QUD is assumed to be, *Which girls went to the party?* The listener has a uniform prior distribution over worlds.

There are three alternative utterances: *all*, *some*, and u_{null}. The semantics for the utterances are given by:

$$[\![all]\!] = \lambda\theta\lambda X\lambda Y[(X \cap \theta) \subset Y] \tag{3.44}$$

$$[\![some]\!] = \lambda\theta\lambda X\lambda Y[\,|(X \cap Y \cap \theta)| > 0] \tag{3.45}$$

The variable θ is the domain for the quantifiers.[10] This domain can be any nonempty subset of the set of girls G. The listener is assumed to have a uniform prior distribution over the quantifier domain. The three quantifiers are assigned cost 1. As in the previous section, the utterance u_{null} is defined to be true in every world, and is much more costly than every other utterance (cost 10). It is included only for technical convenience.

The noise distribution is shown in Figure 3-12; it is a slight generalization of the noise distributions used for previous examples. We set the noise probability $p = 0.01$ The speaker has a binary choice of whether to use stress; if the speaker uses stress, this reduces the noise probability by a factor of 2. As in previous examples, stress has cost 0.1.

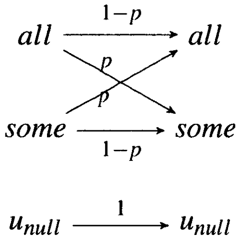


Figure 3-12: The noisy channel for Example (37), when stress is not used.

Strengthening of the universal quantifier

Figure 3-13 shows the predicted effect of stress on utterance interpretation. When the quantifier *all* is stressed, the listener assigns higher probability to all 6 girls having gone to the party; almost all of the listener's probability mass is concentrated on 5 or 6 girls having gone. In contrast, when this utterance is unstressed, the listener infers that fewer girls went to the party.

[10] It is also possible to assign distinct domains to each of the quantifiers. This does not change the qualitative

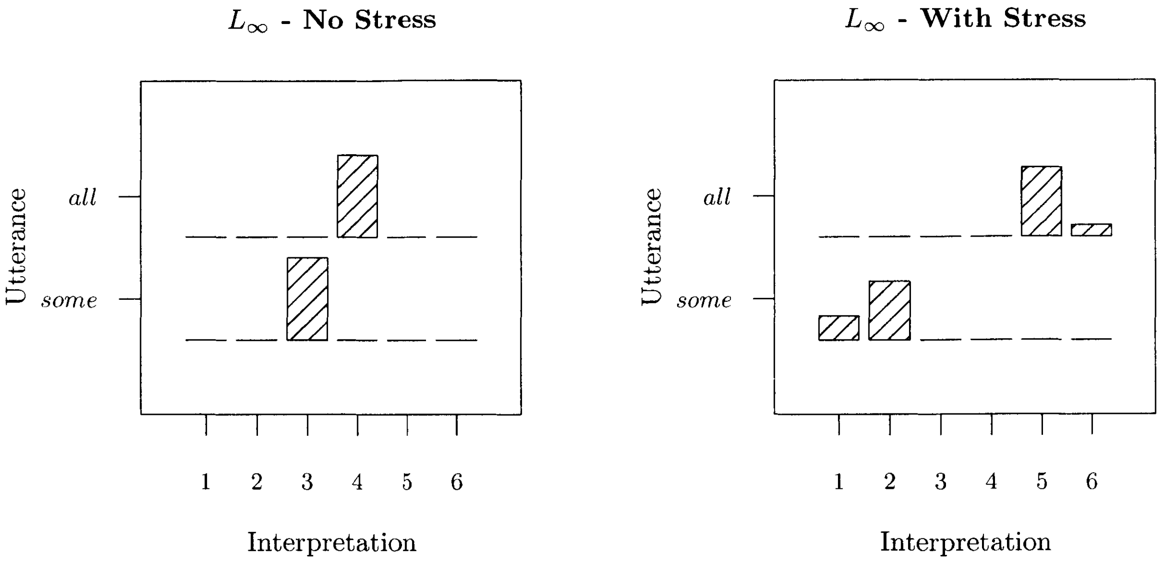


Figure 3-13: The listener's interpretation of the quantifiers without stress (left panel) and with stress. The figures show the marginal probability that the listener assigns to *n* girls having gone to the party, where *n* is indicated on the x-axis.

We will first explain why placing stress on *all* results in the inference that more of the students went to the party. Consider the speaker S_1 who wants to communicate world $\{1,...,6\}$, i.e. that all 6 girls went to the party. For simplicity, suppose that the domain variable θ equals the full domain; the reasoning here will generalize, in a slightly weaker form, to other values for the domain variable as well. In this case, the literal meaning of utterance *all* is maximally informative, as it means that all 6 girls went to the party. The literal meaning of utterance *some*, in contrast, is maximally uninformative in this case, as it only states that at least one out of the 6 girls went to the party — something that does not rule out any worlds. The speaker will choose the utterance *all* with very high probability, as its literal meaning exactly picks out the intended world, and the alternative utterances do not. The speaker will have a strong incentive to ensure that the listener accurately perceives this utterance. If the listener mishears the intended utterance *all* as *some*, then they will gain no information about the speaker's intended meaning; the posterior probability of world

predictions of the model. We assume that there is a single domain primarily to simplify our explanations.

$\{1,...,6\}$ will be same as the prior probability. The speaker will therefore want to decrease the probability that the listener mistakenly hears *some*, and will be likely to place stress on *all*.

Next consider the speaker S_1 who wants to communicate a world in which fewer girls went to the party, e.g. a world in which 4 girls went. Suppose for concreteness that the speaker wants to communicate world $\{1,...,4\}$ (the situation is symmetric for speakers who want to communicate other 4-element worlds). If the domain variable θ contains elements outside of $\{1,...,4\}$, then the speaker will not use the utterance *all*, as the literal meaning of the utterance in this case will be false. We will therefore consider the case in which the quantifier domain equals $\{1,...,4\}$; in this case, the domain is maximally large, subject to the constraint that utterance *all* is literally compatible with the intended world. Our argument generalizes to other, smaller domains as well. The speaker in this case will be likely to use utterance *all*: the utterance will literally communicate that all of the girls in $\{1,...,4\}$ went to the party, while the alternative *some* will not do this.

Relative to the speaker who wants to communicate world $\{1,...,6\}$, however, the speaker who wants to communicate $\{1,...,4\}$ will have weaker incentive to place stress on their utterance. If *all* is not faithfully perceived by the listener, then this is not as much of a problem for this speaker, for two reasons. First, the utterance *all* is not as informative for this speaker as it is for the speaker who wants to communicate world $\{1,...,6\}$. When the speaker wants to communicate $\{1,...,6\}$ (and the domain is sufficiently large), utterance *all* will exactly (or nearly exactly) pick out this world. In contrast, when the speaker wants to communicate world $\{1,...,4\}$, the literal strength of *all* is necessarily limited by its restricted domain. Even when the domain is the maximal one consistent with this world ($\theta = \{1,...,4\}$), it does not provide any information about the individuals outside of the domain (individuals 5 and 6), and therefore does not uniquely pick out the intended world. The second reason that noise is less of a problem for this speaker, is that the alternative utterances are more informative than for the speaker who wants to communicate world $\{1,...,6\}$. Suppose that the listener mishears *all* as *some*, with the quantifier domain $\theta = \{1,...,4\}$. The listener will then infer that at least one element of $\{1,...,4\}$ went to the party. While this clearly is less informative than if the utterance *all* had been accurately transmitted, it nonetheless increases the probability that the listener assigns to world $\{1,...,4\}$, as it rules out the worlds in which only the individuals 5 and 6 went to the party. Thus there are two factors which reduce the speaker's incentive to lower the noise rate. The speaker S_1 who wants to communicate world $\{1,...,4\}$ will be less likely to place stress on utterance *all*.

When the listener L_1 hears that the speaker has placed stress on utterance *all*, they will interpret the use of stress as indicating that a greater number of girls went to the party. The listener L_1 does not know which quantifier domain the speaker S_1 believes is being used for literal interpretation. The listener will therefore interpret the speaker's utterance by averaging over the different possible quantifier domains. On average, the speaker S_1 is more likely to use stress when they want to communicate a world in which a greater number of girls went to the party, e.g. world $\{1,...,6\}$. The listener L_1 will therefore interpret stress as indicating one of these larger cardinality worlds. This interpretation of stress will propagate to the higher-depth speakers and listeners. Higher-order speakers will know that stress on utterance *all* signals that more girls went to the party, and will use stress to communicate this; higher-order listeners know that stress will be used on *all* when more girls have gone to the party, and will interpret stress as conveying this meaning.

Referential interpretation of *some*

We have so far explained the model's predictions for how stress influences the interpretation of utterance *all*. Figure 3-13 also shows the model's predictions for the interpretation of *some*. When *some* is stressed, the listener infers that fewer of the girls went to the party. Stress therefore has opposite effects in the cases of *some* and *all*: rather than signaling an expansion of the set of elements with a certain property, as in the case of *all*, stress on *some* signals a contraction of this set.

It is important to clarify the interpretation of this prediction. Previous work [84] has discussed stress in the context of strengthening scalar implicatures, as in the following contrast:

(41) Some of the students passed the test.

(42) SOME of the students passed the test.

In Example (41), in which the utterance is left unstressed, the scalar term generates a weak implicature: the speaker merely does not know that the scalar alternative *all* is true. The use of stress, as in Example (42), generates a stronger implicature: that the speaker knows that the scalar alternative *all* is false. This type of scalar implicature strengthening is discussed in more detail in Section 3.6. Crucially, the distinction between weak and strong implicatures depends on the possibility of speaker ignorance. If the speaker is assumed to know whether the alternative *all* is true, then the weak implicature (in which the speaker does not know that *all* is true) and the strong implicature (in which the speaker knows that *all* is false) are equivalent. Indeed, standard models of pragmatic reasoning [33, 28, 29] predict that the scalar term *some* should generate the strong implicature, even in the absence of stress, if it is part of the common ground that the speaker is knowledgeable. Thus, it is not clear that these models predict any differences in interpretation between the stressed scalar item and the unstressed one, in the case that the speaker is presumed to be knowledgeable.

The distinctive feature of the prediction in Figure 3-13 is that it applies even given the assumption of speaker knowledgeability. When stress is placed on *some*, it leads to a contraction of the set of individuals being predicated with the property. This prediction was derived given the assumption that the speaker knows which exact world they are in. Standard accounts of pragmatic reasoning do not derive this effect, or comparable ones.

Until now, we have left open the question of whether this prediction is realized by any actual empirical phenomena. The following examples suggest that it may:

(43) Context: Bob is addressing his students before class. He looks at a particular student.
Bob: SOME of you didn't hand me your homework yesterday. [Understood meaning: The student that Bob is looking at did not hand in his homework yesterday.]

(44) Context: Alice is judging a cooking contest. She has a row of pastas from different teams in front of her, and she looks at one.
Alice: SOME of the pastas were overcooked. [Understood meaning: The pasta that Alice is looking at was overcooked.]

In these examples, the quantifier *some* is being used in a *referential* manner: it conveys that a particular individual has some property, e.g. that of not handing in their homework. Though this claim requires further empirical evaluation, this referential reading is most salient when stress is placed on *some*. In the absence of stress, Example (43) most clearly conveys that some but not all of the students handed in their homework, and nothing more, while Example (44) most clearly conveys that some but not all of the pastas were overcooked. The use of stress therefore appears to drive the referential reading.

This reading of *some* is distinct from the normal existential reading of *some*, which posits that at least one individual has some property. It is also not straightforwardly reducible to the strengthened scalar implicature discussed above, which only would imply that not all individuals (in Example (43), not everyone in the class) have the property. The reading is, however, congruent with the predictions of our model. The model predicts that the use of stress on *some* will restrict the predication to a single individual, or a small set of individuals. In the absence of stress, the model predicts that the set of individuals being predicated is expanded, and that the referential reading will be less available. The model therefore correctly predicts that stress will be more likely to signal a referential reading of the quantifier.

Though the model appears to capture some features of stress interpretation in Examples (43) and (44), this claim needs to be qualified. In both scenarios, it appears that the quantifier *some* is being used in order to avoid stating the intended meaning more directly. In Example (43), Bob uses *some* in order to avoid referring to the offending student by name; in Example (44), Alice uses *some* in order to avoid directly naming the only poorly-performing team. In both examples, the speaker's use of the quantifier appear to be driven by politeness. More precisely, they appear to be driven by the speaker wanting to avoid being impolite, by directly criticizing someone in public. Our model does not represent politeness norms, and therefore potentially excludes an important feature driving these inferences. Our claim that the model explains the inferences in Examples (43) and (44) is therefore provisional. Further work is required to determine whether politeness norms are indeed necessary to explain these cases, and, if so, whether the model's predictions are robust to the inclusion of these norms.[11]

We will now explain the predicted effect of stress placement on *some* in the model. Our modeling assumptions remain the same as in Section 3.8.3: the speaker knows how many girls went to the party, and wants to communicate this to the listener. We will first consider the speaker S_1. We will explain why this speaker will be more likely to place stress on *some* when they want to communicate that a smaller number of girls went to the party, rather than a larger number. Suppose first that the speaker wants to communicate world $\{1\}$, i.e. the world in which only individual 1 went to the party (the situation is symmetric for other one-element worlds). The speaker will always choose utterance *some* when the quantifier domain variable θ satisfies: $\{1\} \subsetneq \theta$. In this case, the utterance *some* is literally compatible with world $\{1\}$ (because it literally conveys that at least one element of θ went to the party), and it is the only utterance literally compatible with this world (because the alternative, *all*, would communicate that some individual outside of $\{1\}$ went to the party). The speaker will also have a strong incentive to reduce the noise rate in this case. If the listener mishears *some* as *all*, then they will infer that some element outside of $\{1\}$ went to the party — something which the speaker knows to be false. The speaker who wants to communicate world $\{1\}$ will therefore be likely to place stress on *some*.

The speaker who wants to communicate a world in which more girls went to the party — e.g., world $\{1,2,3\}$ — will be less likely to place stress on *some*. As just noted, the speaker who wants to communicate world $\{1\}$ will have a strong incentive to use *some*, and place stress on it, whenever the domain variable θ is such that $\{1\} \subsetneq \theta$. Consider any value of θ such that $\{1\} \subsetneq \theta \subseteq \{1,2,3\}$. We will argue that the speaker who wants to communicate world $\{1,2,3\}$ will be unlikely to place stress on *some* given these values for the domain variable. Thus, this speaker will be unlikely to place stress on *some* in

[11] An initial hypothesis is that politeness can be modeled as a set of restrictions on what meanings the speaker can literally convey. For example, if it is considered impolite to publicly criticize someone, then this can be modeled as a restriction on the speaker's alternative utterances, which eliminates any utterances which literally express that a particular individual has a negative attribute. In this case, a model which represents politeness would actually coincide with the one presented in this section. A crucial assumption of the model in this section is that the speaker cannot name individual objects; the only alternative utterances are quantifier expressions. If object names are included among the alternative utterances, then the model no longer predicts the availability of the referential reading for *some*. This referential reading disappears due to pragmatic competition from the object names. Thus, it is possible that we have in fact provided an adequate representation of the scenarios in Examples (43) and (44).

certain scenarios in which the speaker in world $\{1\}$ would be likely to use stress. It is straightforward to show that there are no values for θ such that the speaker in world $\{1,2,3\}$ will be *more likely* to use stress than the other speaker. Thus, this speaker will be on average less likely to use stress. To make this argument, suppose that $\{1\} \subsetneq \theta \subseteq \{1,2,3\}$, and that the speaker in world $\{1,2,3\}$ chooses utterance *some*. Suppose that the listener mishears *some* as *all*. In this case, the listener will believe that all of the elements in $\theta \subseteq \{1,2,3\}$ went to the party — something which the speaker actually knows to be true. Thus, there is no problem for the speaker if the listener mishears *some* in this case. As a result, the speaker will not have an incentive to reduce the noise rate, and will be unlikely to use stress.

3.9 Hyperbole

Stress interacts with several types of figurative language. In the current section, we will consider the effect of stress on the interpretation of hyperbolic utterances. Consider a hyperbolic utterance, as in Example (45):

(45) Bob owes me a million dollars.

Under most circumstances, this utterance will communicate two things: that Bob owes the speaker a large amount of money, and that the speaker is unhappy about this. More generally, hyperbolic utterances are typically used for two communicative purposes [81]. First, they are used to communicate information about the state of the world, typically that it has an extreme value along some dimension. Second, they are used to communicate information about the speaker's attitude towards this state, e.g. the speaker's affect.

Stress is often used to amplify the effect of hyperbolic utterances:

(46) Bob owes me A MILLION dollars.

(47) It weighs a TON.

(48) He NEVER says hello.

In each of these examples, the use of stress gives the utterance a more extreme interpretation. In Example (46), stress indicates that Bob owes the speaker an especially large amount of money, and that the speaker is especially unhappy about this. More generally, stress is used to communicate that the state of the world is *especially* extreme along some dimension, and that the speaker has an *especially* strong opinion about this.

The interaction between stress and hyperbolic utterances is not easily modeled within the framework of Alternative Semantics or other standard models of focus interpretation. A first challenge for these accounts is providing a formal semantics and pragmatics for hyperbole. To the best of our knowledge, the only existing formal account of hyperbole is the one presented in [53], which extends the Rational Speech Acts model (and which is discussed in more detail below). The questions of how to properly integrate Alternative Semantics or related frameworks with RSA, or how to develop different formal models of hyperbole, are nontrivial ones, and further research would be necessary to determine how Alternative Semantics interacted with models of hyperbole. There is, however, a more fundamental challenge for Alternative Semantics. As previously discussed, Alternative Semantics proposes that the effects of stress are derived through the following mechanism. When the speaker places stress on part of their utterance, this triggers a set of alternatives to this utterance. The listener uses these alternatives to draw inferences about the speaker's intended message. For structurally-simple declarative sentences, the listener will typically infer that the speaker intended to communicate that the alternatives are false, i.e. they will take the negation of the alternatives. It is not clear how to use this mechanism to derive the effects of stress on the interpretation of hyperbole. In Example (46), the alternatives generated by the placement of stress on *a million* (to a first approximation) will be sentences of the form *Bob owes me X dollars*. Exhaustifying with respect to these alternatives will not derive an amplified interpretation of Example (46). Taking the negation of the alternatives, such *Bob owes me a thousand dollars* or *Bob owes me a billion dollars*, will result in the inference that Bob owes the speaker exactly one million dollars — not the inference that Bob owes the speaker an especially large amount of money, or that the speaker is especially unhappy about Bob. The reasoning here can be generalized: the inferences generated by Alternative Semantics do not appear to have the appropriate form for capturing the effect of stress on hyperbolic utterances.

3.9.1 A model of hyperbolic utterances

In order to explain how the noisy-channel model derives these phenomena, we will present a formalization of Example (46) in this model. This formalization will serve two functions. First, it will demonstrate the principles which allow the model to derive hyperbolic interpretations of utterances — a necessary precursor for explaining the inferences we are interested in. Second, it will demonstrate why the model assigns stress an emphatic interpretation when it is used in hyperbolic utterances. In developing this formalization, we will build on the approach presented in [53]. That work showed to how extend the RSA model in order to explain the interpretation of hyperbolic utterances, but did not address the use or interpretation of stress.

We will use the noisy-channel/QUD model presented in Section 3.3. Our representation of the communicative scenario in Example (46) will be highly simplified in several respects, in order to simplify our explanations and distill the reasoning principles captured by the models. We assume that there are three money states, 10, 100, and 1,000,000, corresponding to the amount of money that Bob owes the speaker.[12] There are two affect states, $\bot$ and $\top$. The affect state $\top$ indicates that the speaker has affect (i.e. that they are unhappy about the amount of money that Bob owes), while $\bot$ indicates that they do not have affect. Each world consists of a money-affect pair $(M = x, A = y)$, where x represents the amount of money M that Bob owes, while y represents the speaker's affect A [78]. The set of worlds is assumed to be maximal with respect to money and affect states, so that there are $3 \cdot 2 = 6$ worlds. The speaker is assumed to know exactly how much Bob owes and what their affective state is. The listener's prior beliefs are assumed to satisfy the following constraints. First, $P(A = \top | M = 10) < P(A = \top | M = 100) < P(A = \top | M = 1{,}000{,}000)$. This captures the intuition that the speaker is more likely to be unhappy when Bob owes greater amounts of money. Second, $P(M = 1{,}000{,}000) < P(M = 100) \leq P(M = 10)$, that is, Bob is less likely to owe a million dollars than he is to owe smaller amounts of money. The particular values of the prior that we use for model simulations are shown below. The qualitative predictions of the model are robust across a wide range of values for the prior,

[12] The qualitative predictions of the model do not change if the set of worlds is made denser, i.e. by including a greater range of worlds between 10 and 1,000,000.

subject to some constraints which are imposed by the noise rate and which are discussed below.

M	$P(M)$	$P(A = \top \mid M)$
10	0.495	0.1
100	0.495	$\frac{1}{3}$
1,000,000	0.01	0.9

Table 3.1: The prior distribution over worlds. The column labeled $P(M)$ provides the marginal probability of each value of M, while $P(A = \top|M)$ provides the conditional probability of affect $\top$.

There are three possible speaker QUDs. First, the speaker may have the QUD q_M, in which case they want to communicate the value of M, i.e. how much money Bob owes them. Second, they may have the affect QUD q_A, in which case they want to communicate the value of A, i.e. their affect state. Finally, they may have the joint QUD $q_{(M,A)}$, in which case they want to communicate the values of both M and A, i.e. which exact world they are in. We assume that the listener has a uniform prior distribution over the speaker's QUD.

The speaker has three alternative utterances: *ten*, *one hundred*, and *one million*. These are given an exact semantics, though the predictions of the model do not differ if they are given a lower-bound semantics:

$$
\begin{aligned}
&[\![ten]\!] = \lambda(M,A).[M = 10] \\
&[\![one\ hundred]\!] = \lambda(M,A).[M = 100] \\
&[\![one\ million]\!] = \lambda(M,A).[M = 1{,}000{,}000]
\end{aligned}
$$

The utterances are assigned cost 1.

The noise distribution is our standard substitution channel, applied to the three alternative utterances, and is shown in Figure 3-14. The noise probability is set to $p = 0.01$. The speaker has the binary choice to use stress or not, and using stress decreases the noise rate by a factor of 2. Stress has cost 0.1 for the speaker.

The model's predictions are shown in Figure 3-15. When the speaker uses utterance *one million* without stress, it is interpreted hyperbolically: the listener assigns high probability

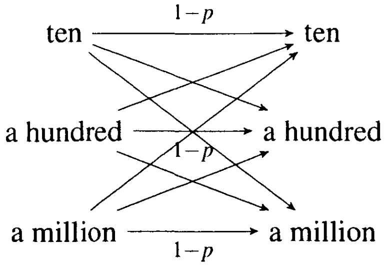


Figure 3-14: The noisy channel for Example (46), when prosodic stress is not used. Utterances are transmitted correctly with probability $1 - p$, and are corrupted with probability p. When an intended utterance is corrupted, the perceived utterance is sampled uniformly from its alternatives.

to Bob owing a large amount of money (though not an implausibly large amount), and to the speaker being unhappy about this. When this utterance is used with stress, the hyperbolic interpretation is amplified: the listener assigns even higher probability to Bob owing a large amount of money, and to the speaker being unhappy.

We will first explain how the model derives a hyperbolic interpretation for the utterance *one million*. In order to simplify this discussion, we will assume that there is no noise, and that stress is unavailable to the speaker (and that this is common knowledge amount the speaker and listener); these assumptions will be relaxed below when we discuss the effects of stress. We start by considering the speaker S_1 who is in world $(100, \top)$, i.e. in which Bob owes a hundred dollars and in which they are unhappy about this. Suppose that this speaker has QUD q_A, so that they want to communicate their affect. The speaker has three alternative utterances available. Neither *ten* nor *one hundred* will effectively communicate the correct answer to their QUD. If they choose *ten*, then the literal listener L_0 will infer that Bob owes ten dollars, and hence, by Table 3.1, that the conditional probability of affect $\top$ is 0.1. Similarly, if they choose *one hundred*, then listener L_0 will infer that Bob owes a hundred dollars, and that the conditional probability of affect $\top$ is $\frac{1}{3}$. The utterance *one million* will be more effective at communicating affect $\top$: if the listener hears *one million*, they will infer that this affect has probability 0.9. The speaker will therefore choose utterance *one million* with high probability if they want to communicate affect $\top$.

Next consider the pragmatic listener L_1 who hears the utterance *one million* (without stress). This listener will consider several possible interpretations of the utterance. The

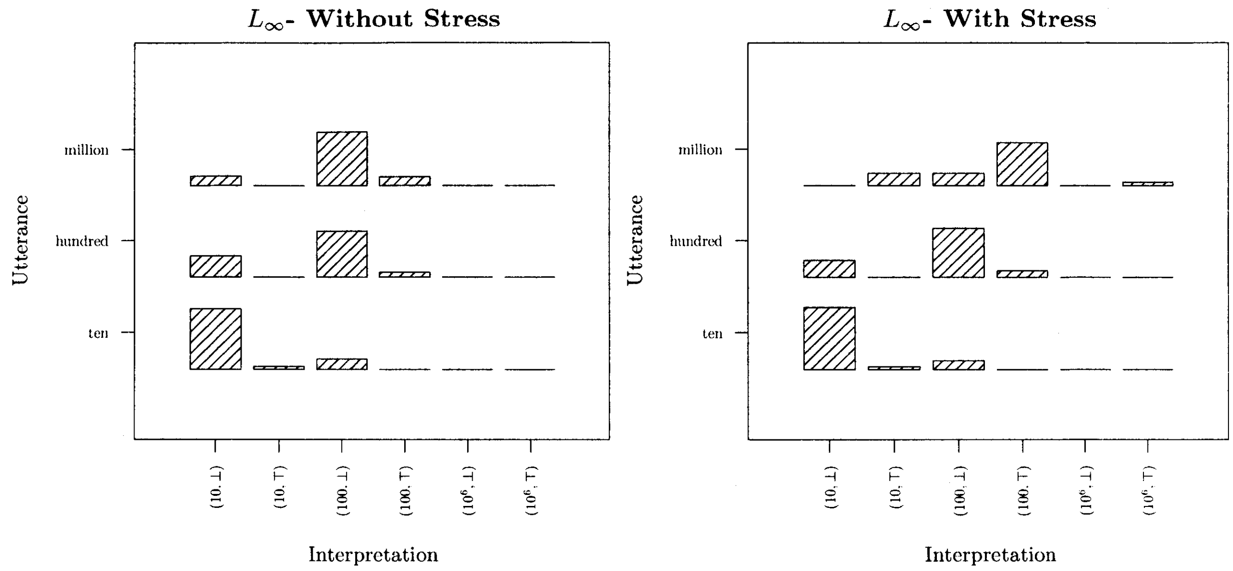


Figure 3-15: The listener's interpretation of the utterances with and without stress. The x-axis indicates the possible money-affect pairs.

speaker would have been likely to choose this utterance if they had wanted to communicate the state $M = 1{,}000{,}000$, or if they had wanted to communicate the worlds $(1{,}000{,}000, \perp)$ or $(1{,}000{,}000, \top)$. However, the state $M = 1{,}000{,}000$, and both of these worlds, are extremely *a priori* unlikely, i.e. Bob is very unlikely to actually owe a million dollars. What is more likely is that the speaker is in the *a priori* probable world $(100, \top)$. As discussed above, the speaker who wants to communicate $\top$ will be likely to choose the utterance *one million*. Thus, the speaker being in world $(100, \top)$ and wanting to communicate their affect provides a plausible explanation for this choice of utterance.

This example illustrates the general principles by which the model derives hyperbolic interpretations. A speaker who wants to communicate high affect can do so effectively by stating that the world is unusually extreme along some dimension. In the current example, worlds in which Bob owes a lot of money are likely to be those in which the speaker is unhappy, and therefore, by literally expressing that Bob owes a lot of money, the speaker will effectively communicate their unhappiness. The speaker will choose utterances which literally express extreme world states (which they know to be false), in order to communicate their affect. The listener knows this, and when they hear an utterance which expresses that the world is surprisingly extreme along the relevant dimension, they will infer that the speaker did not actually want to communicate something extremely implausible. It is more probable that the speaker is in a high-affect world, and that they chose the utterance to communicate this. Typically, the listener will not only gain information about the speaker's affect from hyperbolic utterances, but also information about the state of the world. This is because of the listener's *a priori* knowledge that the state of the world and the speaker's affect are likely to be linked. In the current example, if the speaker is especially unhappy, then it is likely that they are in a world in which Bob owes a lot of money (though less than a million dollars).

Given this explanation of how hyperbolic interpretations are derived in the model, we will now turn to the effects of stress. For this part of the discussion, we will remove our simplifying assumption of noiseless communication. Suppose that the speaker S_1 is in world $(100, \top)$, and has QUD q_A, i.e they want to communicate their affect. For the reasons discussed above, this speaker is likely to choose the utterance *one million*; this utterance is the most effective at communicating that the speaker has affect $\top$. Given the possibility of noise, however, the situation is more complicated than it was above. If the literal listener L_0 hears *one million*, then they have several options for interpreting this utterance. They can interpret this utterance according to its literal meaning, and infer that Bob actually owes a million dollars. This is, however, implausible under the listener's prior distribution. An alternative interpretation is that the speaker's utterance was corrupted by noise, and that the speaker may have intended some other, more literally plausible utterance. If the listener infers that some other utterance was actually intended, then the speaker will likely fail to communicate the intended affect state $\top$; this affect is less likely under the literal interpretation of the alternative utterances. Noise therefore poses a serious problem for the speaker who wants to communicate affect $\top$, and they will have a strong incentive to prevent the listener from inferring that the utterance was corrupted. If the speaker places stress on *one million*, then the listener L_0 will know that the noise rate has been reduced, and will be more likely to infer that *one million* is actually the speaker's intended utterance. Given stress, the listener will be more likely to infer the correct answer to the speaker's QUD, that the speaker is in affect state $\top$. As a result, the speaker who wants to communicate affect state $\top$ will be likely to place stress on *one million*, in order to convince the listener L_0 that this literally implausible utterance was actually intended.

We next consider the listener L_1 who has heard the utterance *one million*, either with or without stress. Due to its *a priori* implausibility, the listener knows that the speaker is unlikely to have intended to communicate that Bob actually owes a million dollars. Rather, it is more likely that the speaker has affect $\top$, and was using the utterance non-literally in order to communicate this. The listener knows, from the reasoning above, that if the utterance is used without stress, then it will be less effective at communicating this affect than if it is used with stress. The listener will therefore be more certain of the speaker's intentions if they hear *one million* with stress: they will infer that the speaker wanted to communicate the affect $\top$, and chose the utterance and stress which are most effective at doing so.[13] Thus, the use of stress provides a stronger signal that the speaker is unhappy.

This example illustrates a general pattern of reasoning which is captured by the model. Hyperbole arises in the model, when the speaker wants to communicate a strong affect, and does so by convincing the listener that the world has an extreme value along some (relevant) dimension. When the world is extreme along some dimension, the listener will infer that the speaker is likely to have a strong attitude towards this. Given the possibility of noise, however, communicating that the world takes on an extreme value is more challenging than it may first appear: if the literal listener hears an utterance which expresses that the world is implausibly extreme along some dimension, then they may infer that the speaker intended some other, more plausible utterance. The speaker will therefore have a strong incentive to reduce the noise rate by placing stress on the utterance. When the pragmatic listener hears stress on the utterance, they will infer that the speaker wanted to communicate a strong affect, and used stress in order to ensure that the listener would not dismiss the literal content of their utterance.

[13] The reasoning here is somewhat subtle. The unstressed utterance will sometimes be explained away by the literal listener, who will interpret it as expressing some alternative, more plausible meaning. As a result, the speaker will sometimes use the unstressed utterance to communicate one of these more plausible meanings. The literal listener is less likely to explain away the stressed utterance, and as a result, the speaker is less likely to use it to communicate one of the alternative meanings. The pragmatic listener knows that the stressed utterance will be more restricted in its use, and is therefore more likely to interpret it as indicating the affect $\top$.

3.10 Association with focus

A central topic in the study of stress interpretation is the interaction between stress placement and so-called*focus-sensitive* lexical items. Consider the following example:

(49) The translator only SPEAKS Chinese.

(50) The translator only speaks CHINESE.

Example (49) communicates that the translator speaks Chinese, but does not write it. Example (50) communicates that the translator speaks Chinese, but does not speak anything else. This contrast may initially appear similar to the one studied in Section 3.5:

(51) The translator SPEAKS Chinese.

(52) The translator speaks CHINESE.

As in Example (49), when stress is placed on *repairs*, Example (51) will communicate that the mechanic repairs engines, and does nothing else to them. As in Example (50), when stress is placed on *engines*, Example (52) will communicate that the mechanic repairs engines and nothing else. It is therefore not clear what role the lexical item *only* is playing in Examples (49) and (50); it may seem initially plausible that it is semantically vacuous, and that stress has the same effect both with and without this lexical item.

This hypothesis can be ruled by by considering the interpretation of the examples in embedded contexts:

(53) If the translator only speaks CHINESE, then we can hire him.

(54) If the translator speaks CHINESE, then we can hire him.

In both of these examples, stress is placed on *Chinese*. If *only* is semantically vacuous, then both examples should be interpreted identically. But consider the conditions under which these utterances can be felicitously used. Suppose that there are three types of translators: those that speak Chinese and French, those that speak Chinese but not French, and those that speak French but not Chinese. Suppose first that the regulations prohibit the hiring of translators who speak French. Example (53) can clearly be felicitously used in this case. Example (54), however, cannot be used in this case: it expresses that the translator can be hired if he speaks *at least* Chinese, which includes translators who also speak French. Thus, the lexical item *only* contributes to the semantic content of the antecedent in Example (53). A similar contrast can be observed when stress is placed on *speaks* rather than *Chinese*:

(55) If the translator only SPEAKS Chinese, then we can hire him.

(56) If the translator SPEAKS Chinese, then we can hire him.

Suppose that there are translators who speak and write Chinese, translators who speak but do not write it, and translators who write but do not speak it. Suppose further that the hiring of translators who write Chinese is prohibited. Example (55) could be felicitously used in this case, while Example (56) could not.

The behavior of *only* in embedded contexts provides evidence that it contributes to utterances' truth conditions. In Example (49), *only* conveys that the translator does not bear any other (relevant) relation to Chinese, e.g. he does not know how to speak Chinese. In Example (55), *only* maintains this semantic force, conveying that the antecedent of the conditional is only satisfied if the translator does not bear any other relation to Chinese. This contrasts with the simple declarative sentence in Example (52), which loses its exhaustive interpretation when embedded in a conditional, as in Example (56).

The lexical item *only* is described as *associating with focus* because its semantic force appears to be sensitive to the placement of stress within an utterance. As already noted, the use of *only* in Examples (53) and (55) changes the semantic content of the antecedents in these utterances. This semantic content is not simply a function of the lexical items used in the antecedent. Though both utterances are lexically identical, they receive different interpretations, owing to the distinct stress placement within them. In Example (53), the placement of stress on *Chinese* indicates that the translator bears the relation of speaking with no other (relevant) language; in Example (55), the placement of stress on *speaks* indicates that the translator bears no other relation with Chinese. The placement of stress thus appears to change the semantic force of *only*.

The term *association with focus* is used to describe this type of semantic dependence between lexical items and stress placement. This terminology is, however, an unfortunately theoretically-loaded way of labeling the phenomenon. *Focus* refers to a theoretical entity which is posited by most theories of stress interpretation. Under these accounts, focus is part of the semantic representation of a sentence, attaching to some phrase within the sentence, while stress is the externalized marker of this representation. This distinction between the semantic representation of focus and its externalization is not an idle one in these accounts: focus location is claimed not to correspond in a fully one-to-one manner with stress location, and various rules and sets of constraints have been proposed in order to explain how focus may percolate from the stress location. Though the details differ widely, the accounts posit that the semantic force of *only* is sensitive to the placement of focus within the utterance, rather than the placement of stress directly. *Association with focus* describes the general type of dependence that may exist between focus placement and the semantic interpretation of a lexical item.

In our account of stress interpretation, focus does not exist as a part of the semantic representation, and therefore association with focus does not properly exist as a phenomenon. This does not, of course, mean that the underlying empirical phenomena do not exist. We simply describe these phenomena in a somewhat less theoretically loaded manner, as the systematic dependence of the semantic interpretation of certain lexical items on the placement of stress. In the current section, we will be trying to explain the dependence between stress placement and the interpretation of *only*.

3.10.1 Association without focus

Our account of the stress-dependence of *only* is a purely pragmatic one. All theories of *only*, to the best of our knowledge, propose that it is semantically underspecified. That is, they propose that the lexical entry for *only* leaves certain aspects of its semantic interpretation underdetermined, and that the full semantic interpretation of *only* can only be resolved in the context of an utterance (or conversation). These theories differ in how they propose the semantic interpretation of *only* is resolved. Most accounts [83, 84, 5, 10, 99] propose a conventionalized relationship between focus and the interpretation of *only*. That is, part of the semantics for *only* (or the rules for semantic composition) specify how the focus position within an utterance should be used to resolve the semantic interpretation of *only*. Our proposal differs from these semantic resolution accounts. We posit that there is no conventionalized association between the meaning of *only* and the placement of stress (or focus). Rather, in the spirit of the account of [80], we propose that the semantic content of *only* is resolved through pragmatic reasoning.

Our semantics for *only* follows a standard free-variable approach. For simplicity, we will consider the interpretation of *only* in the following example; this will illustrate our approach, and our discussion can be straightforwardly generalized to more complex cases.

(57) John only introduced Bob to Alice.

We will treat *only* as a sentential operator. *Only* first of all asserts (or presupposes)[14] the truth of its prejacent, which is in this case *John introduced Bob to Alice* [44, 1, 47, 48]. In addition, it expresses the following condition:

$$\lambda R.\forall p \in R\ (\text{true}(p) \rightarrow p = \mathit{introduced}(j,b,a)) \tag{3.46}$$

Here, R is a free variable which specifies a set of propositions, and $\mathit{introduced}(j,b,a)$ is the proposition that John introduced Bob to Alice. *Only* exhaustifies with respect to the propositions in R. The semantics in Equation 3.46 states that Example (57) is true precisely when every proposition in R other than $\mathit{introduced}(j,b,a)$ is false. *Only* can thus be seen as quantifying over the propositions in R, and asserting that they are false.

We will use the model of semantic free variable resolution from Section 3.8.1 in order to explain how the value of R is resolved. Under this model, inference over semantic free variables is *lifted* to the listener L_1. The literal listener L_0 and speaker S_1 are each parameterized by the values of the semantic free variables, and therefore have no uncertainty about

[14] The distinction between asserted and presupposed content will not be relevant for the discussion in this section.

these values. The listener L_0 interprets utterances according to the literal meaning implied by the assigned values for the free variables. The speaker S_1 chooses utterances in order to be interpreted correctly by the listener L_0, given that this listener will interpret utterances using the speaker's assigned values for the free variables. The pragmatic listener L_1 is *a priori* uncertain about values of the free variables, i.e. about what values L_0 and S_1 believe these variables have. When the listener L_1 hears an utterance, they perform joint inference over the values for the free variables and the other model parameters: the true state of the world, the speaker's QUD, and the speaker's intended utterance.

Given the definitions from Section 3.8.1, we will now present the modeling setup which we use to explain the interpretation of *only* in Example (57). We will assume that each possible world is specified by the set of people who were introduced to each other by John. That is, each world is specified by a set of pairs $\{(x_1,y_1),...,(x_n,y_n)\}$, where the membership of (x_i,y_i) in the set indicates that John introduced x_i to y_i. There are two individuals who can occupy the x role in this relation (i.e. individuals who may have been introduced to someone else), Bob and Charlie, and two other individuals who can occupy the y role (i.e. individuals who may have had someone introduced to them), Alice and Eve. Bob was introduced to some (possibly empty) subset of Alice and Eve, and similarly for Charlie. We only assume that at least one person was introduced to someone. There are thus $2^2 \cdot 2^2 - 1 = 15$ possible worlds. The speaker is assumed to know exactly who was introduced to whom, and the listener has a uniform prior distribution over worlds.

The speaker has two possible types of QUDs. First, the speaker may want to answer a question of the form *Who did John introduce to Y?* Here, Y is someone who may have had someone introduced to them, so Y is either Alice or Eve. This QUD type is denoted by $q_{.Y}$. For example, $q_{.A}$ refers to the question *Who did John introduce to Alice?* Second, the speaker may want to answer *Who did John introduce X to?* Here, X is someone who may have been introduced to someone by John, so X is either Bob or Charlie. We use the notation $q_{X.}$ to denote this QUD type. The questions are given their standard semantics.

The listener has a joint prior probability distribution over worlds and QUDs. Given a world w, the conditional distribution over QUDs is defined as follows. For every world, some subset of the QUDs have non-empty answers in that world. For example, in the world $\{(B,A)\}$, in which Bob was introduced to Alice, and nobody else was introduced to anyone, there are two questions with non-empty answers: $q_{B\cdot}$ (*Who did John introduce Bob to?*) and $q_{\cdot A}$ (*Who did John introduce to Alice?*). For each world, the conditional prior distribution over QUDs is defined to be uniform over the QUDs which satisfy this condition.[15]

The alternatives utterances for the speaker are generated from the following grammar:

$$
\begin{aligned}
&S \rightarrow P:L, \;\; S \rightarrow O:P:L \\
&P \rightarrow b, \;\; P \rightarrow c, \;\; P \rightarrow b \wedge c \\
&L \rightarrow a, \;\; L \rightarrow e, \;\; L \rightarrow a \wedge e \\
&O \rightarrow o
\end{aligned}
$$

The atomic terms b, c, a, and e refer to Bob, Charlie, Alice, and Eve, respectively. The atomic term o indicates that *only* is being used. All of the utterances are either of the form $x : y$, where this indicates that John introduced the individuals in x to the individuals in y, or of the form $o : x : y$, where this indicates that John only introduced the individuals in x to the individuals in y. For example, $o : b : a$ corresponds to the utterance *John only introduced Bob to Alice.* We will refer to the phrase in x as the direct object of the expression, and the phrase in y as in the indirect object. The semantics for *only* is given by Equation 3.46 and the preceding discussion. The utterances are assigned cost in proportion to their length. Each atomic term receives cost 1, and the cost of whole utterances is computed by summing the atomic terms it contains.

A distinctive feature of the model defined in Section 3.8.1 is that the listener L_1 performs inference over the value of semantic free variables. In the current case, there is one semantic free variable, R, which determines the set of propositions that *only* exhaustifies over. In order to specify the model for listener L_1, we need to define this listener's prior distribution over the value of R. The propositions in R will be drawn from the following

[15] As in Section 3.5.1, we exclude QUDs which do not satisfy this condition merely in order to simplify the example. If we were to include QUDs with empty answers, then the speaker could only express these answers by using utterances containing negation (e.g. *John did not introduce anyone to Alice*). By excluding these QUDs, we can considerably simplify the set of alternative utterances that need to be considered.

set:

$$\mathcal{R} = \{introduced(j,b,a),\ introduced(j,b,e),$$
$$introduced(j,c,a),\ introduced(j,c,e)\}$$

The proposition $introduced(j,b,a)$ expresses *John introduced Bob to Alice*, and similarly for the other propositions. These propositions correspond to the atomic utterances generated by the grammar above. That is, for each atomic utterance of the form $x : y$, the proposition expressed by this utterance is contained in $\mathcal{R}$. We assume that the variable R is such that $R \in 2^{\mathcal{R}}$. The listener's prior distribution is uniform over the members of $2^{\mathcal{R}}$. Importantly, this implies that the listener has no prior information about whether *only* will provide information about Bob rather than Charlie, or about Alice rather than Eve.

As in Section 3.8.1, the speaker has three available choices of prosodic stress, $\perp$, *Direct*, and *Indirect*. For an utterance $x : y$ or $o : x : y$, the stress $\perp$ indicates that no stress was used, *Direct* indicates that stress has been placed on the direct object x, and *Indirect* indicates that it has been placed on the indirect object y. For simplicity, we assume that stress cannot be placed on *only*. The use of stress *Direct* decreases the probability of noise on x by a factor of 2, and similarly for stress *Indirect*.

The noise distribution is shown in Figure 3-16. The atomic terms in the direct object can be confused for each other, and similarly for the atomic terms in the indirect object. As in previous examples, we assume that conjunctions are accurately perceived. When the speaker uses *only* in the utterance, indicated by o, the listener is assumed to perceive this accurately. Noise is sampled independently for the direct object and indirect object.

Figures 3-17-3-19 show the predicted interpretation of the alternative utterances. When the speaker uses *only*, with stress placed on the direct object, the listener interprets this by exhaustifying over the direct object. For example, when utterance $o : b : a$ receives stress on b, it is interpreted as indicating that Bob, and nobody else, was introduced to Alice. In contrast, when stress is placed on the indirect object, the listener exhaustifies over the indirect object. When utterance $o : b : a$ receives stress on a, it is interpreted as indicating that Bob was introduced to Alice and nobody else.

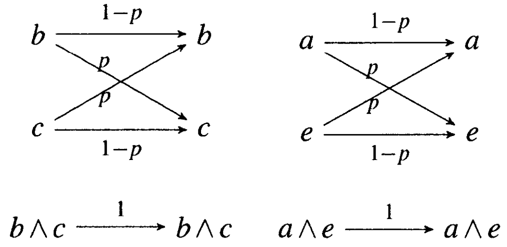


Figure 3-16: Each utterance is composed of a direct object and an indirect object. The noise distribution for indirect objects is shown on the left, while the noise distribution for direct objects is shown on the right. We assume that noise applies independently to the two phrases of the utterance. We assume that *only*, when it is used, is always perceived correctly.

To explain how the model derivations work, we will first consider the speaker S_1. This speaker is parameterized by three variables: the world that they are in, their QUD, and the value for the free variable R, which they believe is being used for literal interpretation. The interaction between these three types of variables makes this scenario especially complicated. We will therefore break our argument down into several parts.

We will consider the speaker with QUD $q._A$, i.e. the speaker who wants to answer the question, *Who did John introduce to Alice?* (The situation for the other three QUDs is symmetric.) There are three possible answers to this QUD: *Bob*, *Charlie*, or *Bob and Charlie*.[16] The first two answers are symmetric, and we therefore can restrict our attention to speakers who want to communicate either the answer *Bob* or the answer *Bob and Charlie*. We will argue that there are two asymmetries between these speakers. First, the speaker who wants to communicate the answer *Bob* is more likely to use *only* than the speaker who wants to communicate the answer *Bob and Charlie*. Second, the speaker who wants to communicate the answer *Bob* is more likely to use the stress *Direct* than the stress *Indirect*.

We will next consider the influence of the semantic free variable R on the speaker's choices. Again, we restrict our attention to the speaker with QUD $q._A$ who wants to communicate the answer *Bob*. There are 16 possible values of the free variable R, but we will focus on two of these values: $R = \{introduced(j,c,a)\}$ and $R = \{introduced(j,b,e)\}$. These settings for R will illustrate a critical asymmetry predicted by the model. We will ex-

[16] These denote exhaustive answers to the question. For example, the answer *Bob* indicates that Bob but not Charlie was introduced to Alice.

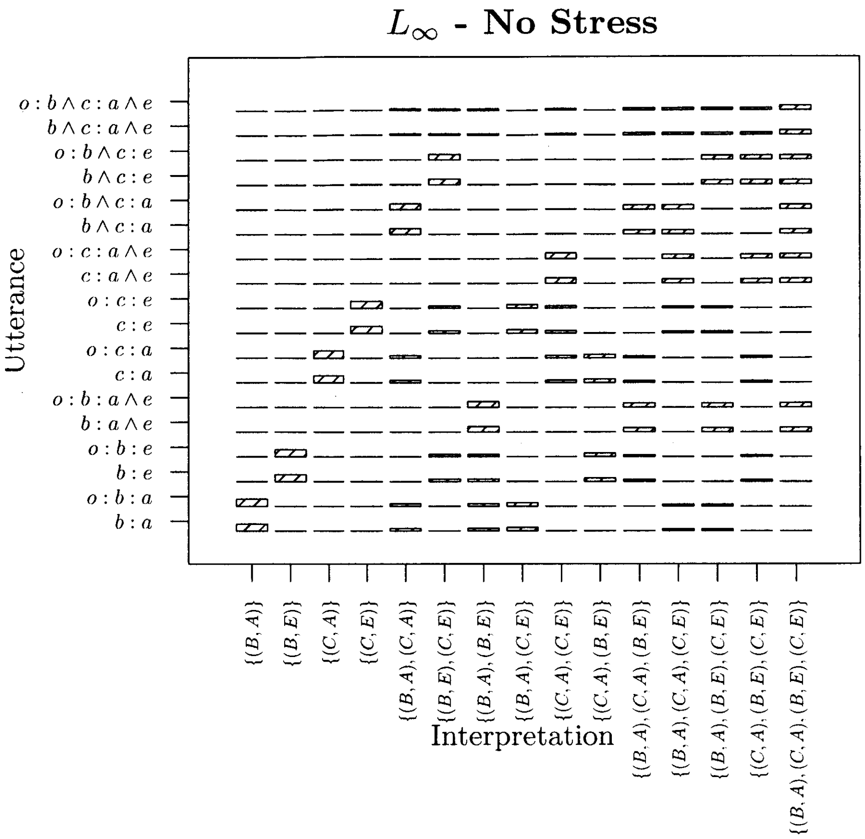


Figure 3-17: The listener's interpretation of the utterances without stress.

plain why the speaker will be more likely to use stress, and in particular to use stress *Direct*, when $R = \{introduced(j,c,a)\}$ than when $R = \{introduced(j,b,e)\}$. This asymmetry will mean that the placement of stress will provide information about the propositions that *only* exhaustifies over.

We will now explain the asymmetry between the speaker S_1 who wants to communicate the answer *Bob* and the speaker who wants to communicate the answer *Bob and Charlie* (given QUD $q_{.A}$). We will first explain why the speaker who wants to communicate *Bob* will be more likely to use *only* in their utterance. Suppose that the speaker wants to communicate *Bob*. This speaker will almost always choose utterance $b : a$ (*John introduced Bob to Alice*) or utterance $o : b : a$ (*John only introduced Bob to Alice*), as no other utterances provide more literal information about the speaker's intended answer to the QUD (and most provide less). For every value of the free variable R, the utterance $o : b : a$ is literally

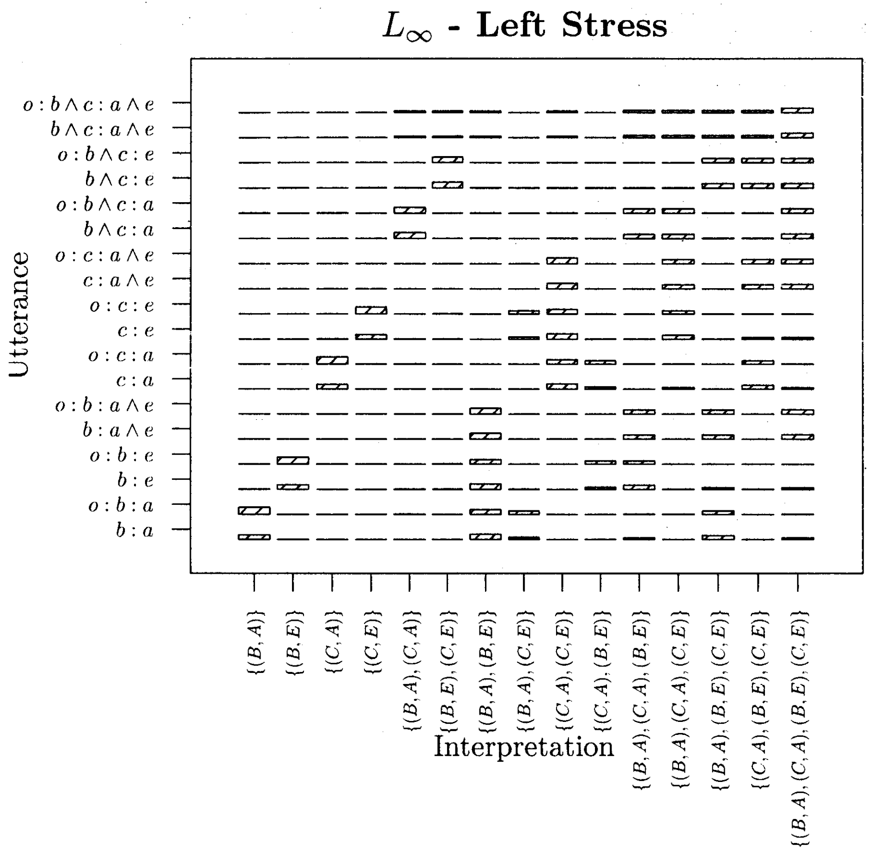


Figure 3-18: The listener's interpretation of the utterances with stress *Direct*.

compatible with the speaker's intended answer to the QUD. When $introduced(j,c,a) \in R$, the utterance $o:b:a$ will communicate the answer to the QUD exactly: it will imply that Bob was introduced to Alice, and that Charlie was not. When $introduced(j,c,a) \notin R$, the utterance $o:b:a$ will not be any more informative for this speaker than $b:a$, but it will still be literally consistent with the intended answer to the QUD. Thus, the speaker who wants to communicate the answer *Bob* to the QUD $q._A$ will be relatively likely to use utterance $o:b:a$. Next consider the speaker with this QUD who wants to communicate the answer *Bob and Charlie*. The utterance $b\wedge c:a$ will be maximally informative for this speaker, but it is somewhat costly due to the conjunction. As a result, this speaker will also sometimes choose utterances $b:a$ and $c:a$, which communicate partial answers to the QUD. This speaker will be unlikely to choose utterances containing *only*, i.e. $o:b:a$ or $o:c:a$. Depending on the value of the free variable R, the utterance $o:b:a$ (and, by

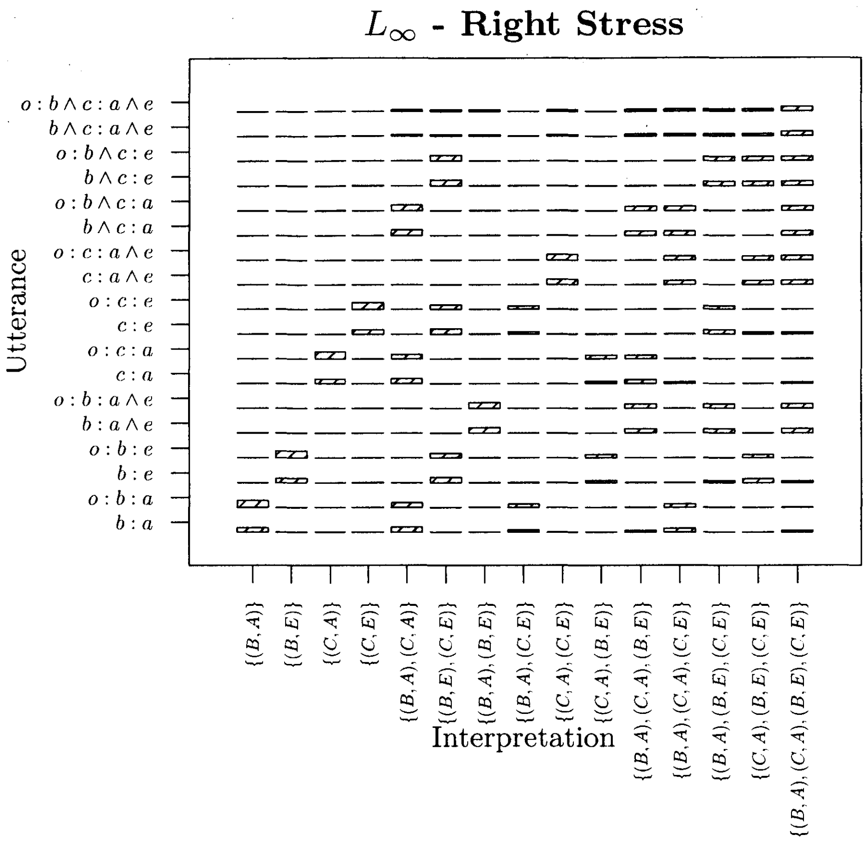


Figure 3-19: The listener's interpretation of the utterances with stress *Indirect*.

symmetry, $o:c:a$) is sometimes inconsistent with the speaker's intended answer to the QUD, and is never more informative for this speaker than the cheaper utterance $b:a$. If $introduced(j,c,a) \in R$, then $o:b:a$ will literally convey that Bob but not Charlie was introduced to Alice, something which is incompatible with the intended answer to the QUD. If $introduced(j,c,a) \in R$, then $o:b:a$ will communicate the same information about the QUD as the simpler utterance $b:a$. The speaker who wants to communicate the answer *Bob and Charlie* to the QUD $q._A$ will therefore be unlikely to use *only* in their utterance.

We will next explain a second feature of the speaker who wants to communicate *Bob* as an answer to the QUD $q._A$: why this speaker will be more likely to use stress *Direct* (i.e. place stress on the direct object) than stress *Indirect*. The reasoning for this point is nearly identical to the reasoning in Section 3.5.2. As discussed above, the speaker who wants to communicate the answer *Bob* is likely to choose either utterance $b:a$ or utterance

$o : b : a$. This speaker will decide whether, and where, to place stress by considering the consequence of having the different parts of their utterance misheard by the listener. Suppose first that the listener L_0 mishears the direct object b, and instead hears c. In this case, whether the speaker intended utterance $b : a$ or $o : b : a$, the listener will believe that Charlie was introduced to Alice, something which is incompatible with their intended answer to the QUD. The speaker will therefore have a strong incentive to decrease the noise rate on the direct object. Suppose instead that the listener mishears the indirect object a, and instead hears e. If the speaker's intended utterance was $b : a$, then this is not a large problem for the speaker. The listener will believe that Bob was introduced to Eve. While this is not something that the speaker wanted to communicate, it is nonetheless compatible with the intended meaning that Bob but not Charlie was introduced to Alice. The speaker will therefore not have a strong incentive to reduce the noise rate on the indirect object of $b : a$. Next consider the case in which the speaker intended utterance $o : b : a$. Whether the speaker has a strong incentive to reduce the noise rate on the indirect object a will depend on the value of the free variable R. If $introduced(j,b,a) \in R$, then if the listener mistakenly hears $o : b : e$, then the listener will conclude that $introduced(j,b,a)$ is false, i.e. that Bob was not introduced to Alice. This is incompatible with the intended answer to the QUD, and therefore the speaker will want to reduce the noise rate on the indirect a in this case. But consider instead the case in which $introduced(j,b,a) \notin R$. If the listener mishears $o : b : a$ as $o : b : e$, then the speaker will not have communicated that Bob was introduced to Alice, but the listener's beliefs will not be incompatible with Bob having been introduced to Alice (or with Charlie having not been introduced to Alice). As a result, the speaker will not have a strong incentive to reduce the noise rate on the indirect object in this case. Thus, while the speaker will always have a strong incentive to reduce the noise rate on the direct object, they will only sometimes have such an incentive to reduce the noise rate on the indirect object. The speaker will, on average, be more likely to use stress *Direct* than *Indirect*, for either utterance $b : a$ or $o : b : a$.

We will now explain the final relevant feature of the speaker S_1's behavior: that given QUD $q._A$ with answer *Bob*, this speaker is more likely to use stress when the free variable $R = \{introduced(j,c,a)\}$ than when $R = \{introduced(j,b,e)\}$. As noted above, this speaker will choose either utterance $b : a$ or $o : b : a$ with high probability. The value of the free variable R is only relevant for utterance $o : b : a$, so we will restrict our attention to this utterance. Suppose first that $R = \{introduced(j,c,a)\}$. In this case, the utterance $o : b : a$ is maximally informative for the speaker: the utterance communicates that Bob was introduced to Alice, but that Charlie was not, which is exactly what the speaker wants to communicate. If the listener mishears the direct object b as c, then the listener will no longer draw either of these inferences. The speaker will therefore have an especially strong incentive to decrease the noise rate in this case: transmitting the utterance accurately means communicating precisely the correct answer to the listener. Contrast this with the case that $R = \{introduced(j,b,e)\}$. When the listener hears $o : b : a$ in this case, they will infer that Bob was introduced to Alice but not Eve. This will partially communicate the speaker's answer to the QUD, but not fully, as it does not imply that Charlie was not introduced to Alice. If the listener mishears the direct object b as c, then this is still a problem for the speaker, as it will communicate something which the speaker knows to be false (just as in the previous case). However, the payoff from accurately transmitting $o : b : a$ is not as great when $R = \{introduced(j,b,e)\}$ as when $R = \{introduced(j,c,a)\}$, and hence the listener will have less incentive to reduce the noise rate on the direct object when $R = \{introduced(j,b,e)\}$. The speaker with $R = \{introduced(j,c,a)\}$ will therefore be more likely to choose stress *Direct*.

Using these facts about the speaker S_1, we will now explain the listener L_1's interpretation of utterance $o : b : a$ when the direct object b receives stress. In particular, we will explain why the listener interprets this utterance as indicating that Charlie was not introduced to Alice, as opposed to indicating that Bob was not introduced to Eve; this is the basic asymmetry which is labeled by *association with focus*. Ignoring stress, for the moment, the first argument above established the following: the speaker with QUD $q_{\cdot A}$, who wants to communicate the answer *Bob*, will be more likely to choose utterance $o : b : a$ than the speaker with this QUD who wants to communicate the answer *Bob and Charlie*. A symmetric argument establishes that the speaker with QUD $q_{B\cdot}$, who wants to communicate the answer *Alice*, will also be more likely to use this utterance than the speaker with this QUD who wants to communicate the answer *Alice and Eve*. As a result, when the listener L_1 hears utterance $o : b : a$, they will conclude that either the speaker wants to communicate the answer *Bob* to QUD $q_{\cdot A}$, or they want to communicate the answer *Alice* to the QUD $q_{B\cdot}$. The placement of stress on the direct object b with break the symmetry between these interpretations. By the argument above, the speaker S_1 is more likely to place stress on the direct object than on the indirect object given QUD $q_{\cdot A}$. A symmetric argument shows that they are more likely to place stress on the indirect object given QUD $q_{B\cdot}$. Thus, given utterance $o : b : a$ with stress on the direct object, the listener will infer that it is more likely that the speaker wanted to communicate the answer *Bob* to QUD $q_{\cdot A}$, than that the speaker wanted to communicate the answer *Alice* to the QUD $q_{B\cdot}$.

We have so far explained the listener's inferences about the state of the world and the speaker's QUD. The listener draws a further type of inference in this case: that the literal meaning of $o : b : a$, as determined by the semantic free variable R, is more likely to exhaustify over the proposition $introduced(j,c,a)$ than over the proposition $introduced(j,b,e)$. That is, the utterance is more likely to literally convey that Charlie was not introduced to Alice, than that Bob was not introduced to Eve. Given the utterance $o : b : a$ with stress on the direct object, the listener assigns highest probability to the speaker who wants to communicate the answer *Bob* to QUD $q_{\cdot A}$. We argued above that the speaker S_1 with this goal will be more likely to place stress on the direct object, when $R = \{introduced(j,c,a)\}$ than when $R = \{introduced(j,b,e)\}$. As a result, the listener will assign higher probability to $R = \{introduced(j,c,a)\}$ than to $R = \{introduced(j,b,e)\}$ when they hear stress on the direct object.

Chapter 4

Conclusion

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

4.1 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 2.3.3 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 work. 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 work).

If these predictions are correct it would suggest a unification of disambiguation and the richer implicature phenomena considered in this work. 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.

4.2 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*.[1] The examples we have considered in this work have been restricted to have only one token of each word, and hence are insensitive to the granularity of refinement.

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, i.e. 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 2.3.4, 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*, i.e. 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, pos-

[1] 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.

sibly all of them." Suppose both instances of "some" are assigned the refinement *at least some*. The listener will still draw the specificity implicature (i.e. 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.

4.3 The flexibility of lexical refinement

In section 2.3.3, 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 2.3.4 and 2.3.5, 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 work 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 2.3.4. 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 A.4, 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 [53, 52, 61] 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.[2]

[2] Lexical uncertainty, as first proposed in **(author?)** [7], assumes that the lexicon fully specifies the se-

These semantic free variables are used to assign underspecified semantic content to utterances. This is done in the usual manner [66, 72]: 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 [60]. 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 work.

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 work. 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 important implication of the free-variable interpretation of lexical uncertainty is that any case previously identified as containing semantic underspecification potentially supports the kinds of complex pragmatic interactions described here. That is, our formalization of pragmatic inference formalizes the pragmatic resolution of contextual variables that have been used since the dawn of compositional semantics [66, 72]. Further research will be needed to determine if this is the right theory of inference for all free variables.

4.4 Psychological implications of the models

The models presented in this work 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 [70], 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

mantic content of each lexical item. The formulation in terms of semantic free variables was proposed in **(author?)** [60].

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, i.e. 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 work 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 [36, 22, 30]. 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 [43, 13].

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 [75, 40, 56].

4.5 Utterance cost and complexity

The notion of utterance cost plays an important role in the explanations of a number of phenomena discussed in this work. 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 [37]: "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 [45]. 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 work, 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 [62]: 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, i.e. 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.

4.6 Embedded implicatures

[20] 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 [14]:

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

(2) 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 (1) conveys information about the temporal ordering of events: the king first died of a heart attack, and a republic was declared subsequently. [37] 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 (2) presents a problem for this analysis. In this example, the sentence in (1) 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 work, 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 (2) to this case.

Relevance theorists such as [93] and [14] (though see [62] for related suggestions) propose an alternative account of semantic and pragmatic content, which can explain cases like Example (2). Under this account, the semantic content of the embedded phrase in Example (2) is underdetermined, i.e. its propositional content is unfixed prior to its use in the sentence. Following the assertion of Example (2), 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 (2) 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 [16]. 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 [16]. 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 [15, 24]. 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 [88, 86]. 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.

4.7 Conclusion

In this work we have explored a series of probabilistic models of pragmatic inference. The initial Rational Speech Acts model [36] 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.

In the second part of this work, we demonstrated the surprising power of noise in pragmatic reasoning. We have argued that the interpretive effects of stress result from the strategic exploitation of noise, which emerges naturally when noisy channel and probabilistic pragmatics models are integrated. This approach was used to explain a number of phenomena related to the interpretation of stress. We first demonstrated that our model derives a number of standard stress-related phenomena: exhaustive interpretations, scalar implicature strengthening, the association between stress and disagreement, and the interpretation of the focus-sensitive adverb *only*. We then showed that it can account for several phenomena which are, to the best of our knowledge, outside of the scope of previous accounts of stress interpretation: the effects of stress on quantifier domain inferences, the intensification of gradable adjective interpretation, and the strengthening of hyperbolic utterances.

Appendix A

Pragmatic Reasoning through Semantic Inference

A.1 Experimental validation of ignorance implicature

Here we will describe an experimental evaluation of the linguistic judgments discussed in Section 2.3.6. For ease of exposition, we will reproduce the examples from that section here:

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

(2) Some of the students passed the test.

The experiment evaluated two claims about the interpretation of example (1). The first claim is that while example (2) implicates that not all of the students passed the test, example (1) 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.1.1 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."

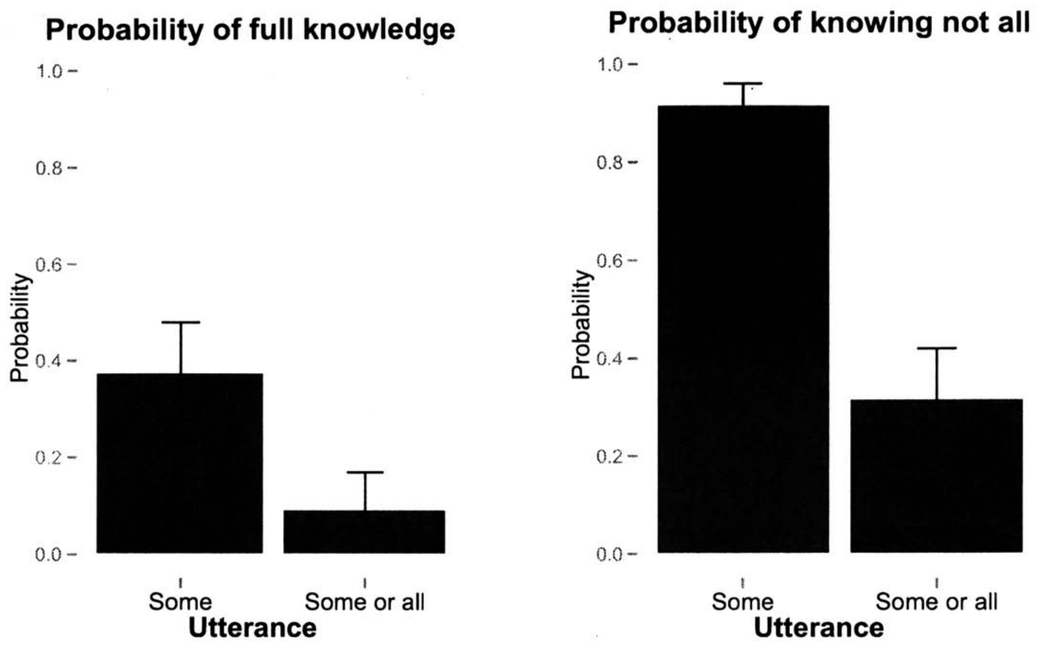
The name of the speaker (e.g. "Laura") and the type of object being observed (e.g. checks inside letters) were varied between items. The speaker's utterance was varied within items, giving two conditions for each item, "Some" and "Some or all." Each participant was shown every item in a randomly assigned 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) $P(\mathbf{A} = 10)$ 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 A-1: Interpretation of the two speaker utterances.

A.1.2 Results

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 A-1a.

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.[1] 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 A-1b.

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.

A.2 Two incorrect definitions of lexical uncertainty

In Section 2.3.3, we defined lexical uncertainty, and in Section 2.3.5, we used this technique to derive M-implicatures. The definition of lexical uncertainty contains several subtle assumptions about the speaker's and listener's knowledge of the lexicon. In this section, we will examine these assumptions in more detail, and will demonstrate that two alternative definitions which violate these assumptions fail to derive M-implicatures.

Consider the definition of lexical uncertainty in Equations 2.24, 2.26, and 2.27. 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

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

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 2.1. The literal listener is now defined as being uncertain about which lexicon to use for interpreting utterances:

$$L_0^{unc}(o,w|u) \propto \sum_{\mathcal{L}'} P(\mathcal{L}')P(o,w)\mathcal{L}'(u,w) \tag{A.1}$$

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 2.3.3.

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 A.1. Define the lexicon $\mathcal{L}_P$ as follows:

$$\mathcal{L}_P(u,w) = \sum_{\mathcal{L}} P(\mathcal{L})\mathcal{L}(u,w) \tag{A.2}$$

Then it follows from equation A.1 that:

$$\begin{aligned} L_0^{unc}(o,w|u) &\propto \sum_{\mathcal{L}'} P(\mathcal{L}')P(o,w)\mathcal{L}'(u,w) && \text{(A.3)}\\ &= P(o,w)\sum_{\mathcal{L}'} P(\mathcal{L}')\mathcal{L}'(u,w) && \text{(A.4)}\\ &= P(o,w)\mathcal{L}_P(u,w) && \text{(A.5)}\\ &\propto L_0(o,w|u,\mathcal{L}_P) && \text{(A.6)} \end{aligned}$$

The last line follows by noting that this is identical to the definition of the literal listener in equation 2.2. Because both $L_0^{unc}(\cdot|u)$ and $L_0(\cdot|u,\mathcal{L}_P)$ define distributions, it follows that

$\mathcal{L}_0^{unc}(\cdot|u) = \mathcal{L}_0(\cdot|u, \mathcal{L}_P)$. □

This lemma shows that the literal listener in the $\mathcal{L}_0$-uncertainty model can be equivalently defined as a literal listener who is certain that the lexicon is $\mathcal{L}_P$. The listener $\mathcal{L}_0$ in the new model is therefore equivalent to a listener $\mathcal{L}_0$ in the rational speech acts model. Because the $\mathcal{L}_0$-uncertainty model is identical to the rational speech acts model for all agents other than $\mathcal{L}_0$, it follows that the $\mathcal{L}_0$-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 $\mathcal{L}_0(\cdot|u, \mathcal{L}_P) = \mathcal{L}_0(\cdot|u', \mathcal{L}_P)$.*

Proof. Let Λ be the set of lexica as defined in Section 2.3.3, 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 $\mathcal{L}_S$, it follows that $f(\mathcal{L})$ is an admissible lexicon iff $\mathcal{L}$ is admissible. Furthermore, because $P(\mathcal{L})$ is the maximum entropy distribution over admissible lexica, $P(\mathcal{L}) = P(f(\mathcal{L}))$.

Given this bijection f,

$$\begin{aligned}
\mathcal{L}_0(o,w|u,\mathcal{L}_P) &\propto P(o,w)\mathcal{L}_P(u,w) && \text{(A.7)}\\
&= P(o,w)\sum_{\mathcal{L}'} P(\mathcal{L}')\mathcal{L}'(u,w) && \text{(A.8)}\\
&= P(o,w)\sum_{\mathcal{L}'} P(f(\mathcal{L}'))f(\mathcal{L}')(u',w) && \text{(A.9)}\\
&= P(o,w)\mathcal{L}_P(u',w) && \text{(A.10)}\\
&\propto \mathcal{L}_0(o,w|u',\mathcal{L}_P) && \text{(A.11)}
\end{aligned}$$

Equality between $\mathcal{L}_0(\cdot|u,\mathcal{L}_P)$ and $\mathcal{L}_0(\cdot|u',\mathcal{L}_P)$ follows from the fact that both define probability distributions. □

These two lemmas have established that the $\mathcal{L}_0$-uncertainty model is an instance of the rational speech acts model, and that the listener $\mathcal{L}_0$ 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:

$$U_1(u|o) = \mathbb{E}_{P_o} \log \frac{1}{L_a(\cdot|u)} - c(u) \tag{A.12}$$

where L_a is defined by:

$$L_a(\cdot|u) = \sum_{\mathcal{L}} P(\mathcal{L}) L_0(\cdot|u, \mathcal{L}) \tag{A.13}$$

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 2.3.3.

Lemma 5. *Let utterances u, u' be assigned identical interpretations by the semantic lexicon $\mathcal{L}_S$. Then, as defined by equation A.13, $L_a(\cdot|u) = L_a(\cdot|u')$.*

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

$$\begin{aligned} L_a(o,w|u) &= \sum_{\mathcal{L}} P(\mathcal{L}) L_0(o,w|u,\mathcal{L}) && \text{(A.14)} \\ &= \sum_{\mathcal{L}} P(\mathcal{L}) \frac{P(o,w)\mathcal{L}(u,w)}{Z_{u,\mathcal{L}}} && \text{(A.15)} \\ &= \sum_{\mathcal{L}} P(f(\mathcal{L})) \frac{P(o,w) f(\mathcal{L})(u',w)}{Z_{u',f(\mathcal{L})}} && \text{(A.16)} \\ &= \sum_{\mathcal{L}} P(f(\mathcal{L})) L_0(o,w|u',f(\mathcal{L})) && \text{(A.17)} \\ &= L_a(o,w|u') && \text{(A.18)} \end{aligned}$$

The term $Z_{u,\mathcal{L}}$ is the normalizing constant for the distribution $L_0(\cdot|u,\mathcal{L})$, and the equality

$Z_{u,\mathcal{L}} = Z_{u',f(\mathcal{L})}$ follows from the fact that $\mathcal{L}(u,w) = f(\mathcal{L})(u',w)$ for all lexica $\mathcal{L}$. □

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

A.3 Refined lexica for non-convex disjunctions

For reference, Figure A-2 depicts the 21 refined lexica in our formalization of Example (11) in Section 2.4.7.

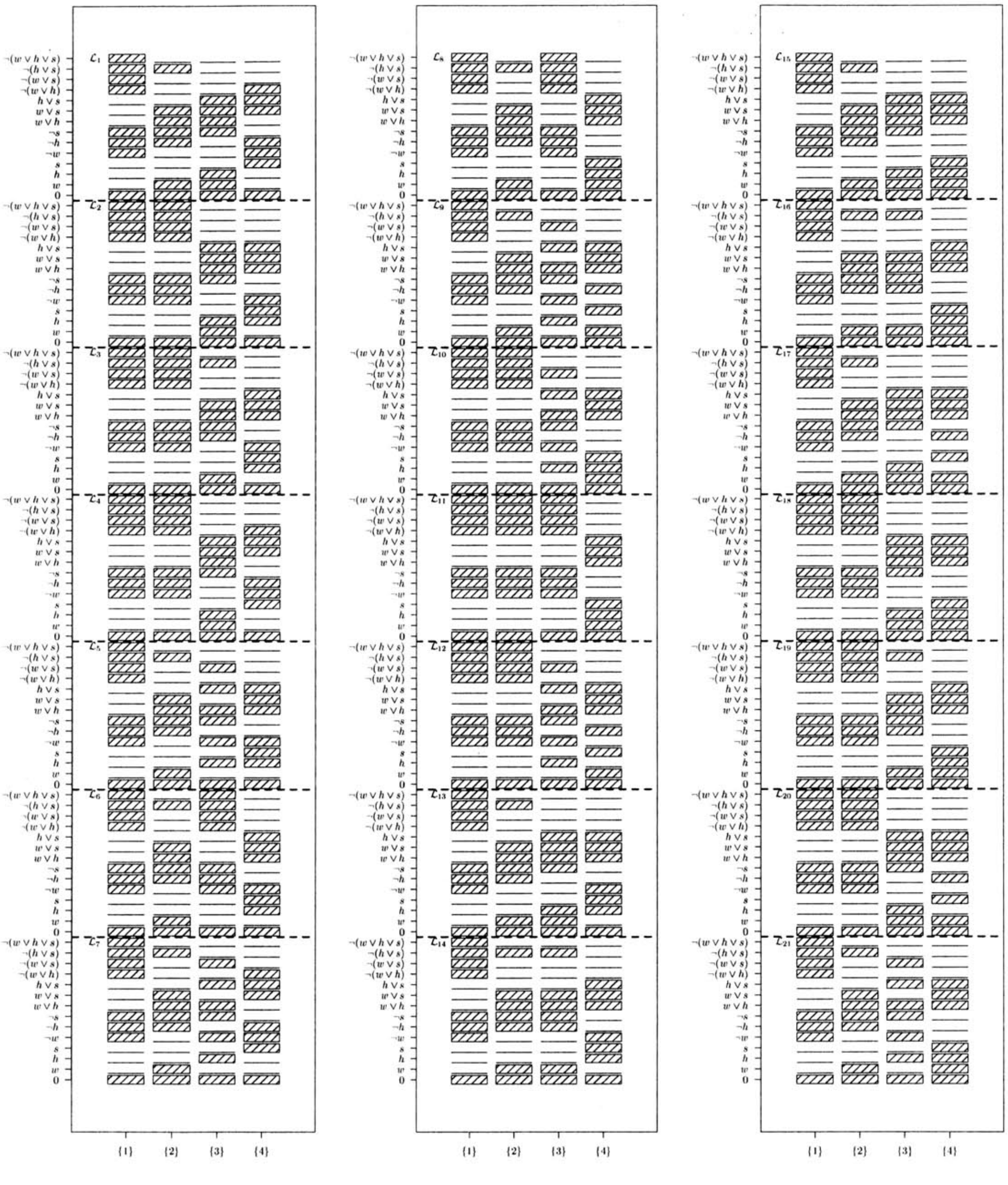


Figure A-2: The 21 refined lexica for formalization of Example (11) in Section 2.4.7.

A.4 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 $(x,y) \in \mathrm{X} \times \mathrm{Y}$.[2] Suppose that the semantic content of each utterance u conveys no information about the dimension Y, i.e. 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,

$$\begin{aligned} L_1((x,y),o|u) &\propto P(o,(x,y)) \sum_{\mathcal{L}} P(\mathcal{L}) S_1(u|o,\mathcal{L}) && \text{(A.19)} \\ &= P(o)P((x,y)|o) \sum_{\mathcal{L}} P(\mathcal{L}) S_1(u|o,\mathcal{L}) && \text{(A.20)} \\ &= P(o)P(\mathrm{X}=x|o)P(\mathrm{Y}=y) \sum_{\mathcal{L}} P(\mathcal{L}) S_1(u|o,\mathcal{L}) && \text{(A.21)} \\ &= P(\mathrm{Y}=y)F(x,o,u) && \text{(A.22)} \end{aligned}$$

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

$$F(x,o,u) = P(o)P(\mathrm{X}=x|o) \sum_{\mathcal{L}} P(\mathcal{L}) S_1(u|o,\mathcal{L}) \qquad \text{(A.23)}$$

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

It follows that

$$
\begin{aligned}
L_1(\mathrm{Y}=y|u) &= \frac{P(\mathrm{Y}=y)\sum_{x,o,u} F(x,o,u)}{\sum_y' P(\mathrm{Y}=y')\sum_{x,o,u} F(x,o,u)} && \text{(A.24)} \\
&= \frac{P(\mathrm{Y}=y)\sum_{x,o,u} F(x,o,u)}{\sum_{x,o,u} F(x,o,u) \cdot \sum_y' P(\mathrm{Y}=y')} && \text{(A.25)} \\
&= \frac{P(\mathrm{Y}=y)}{\sum_y' P(\mathrm{Y}=y')} && \text{(A.26)} \\
&= P(\mathrm{Y}=y) && \text{(A.27)}
\end{aligned}
$$

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.

Appendix B

The Strategic Use of Noise in Pragmatic Reasoning

B.1 Worked example

B.1.1 Distribution over intended utterances

Using the model specification from Section 3.4, we will now illustrate how the listener and speaker computations work, and how particular QUDs get associated with prosodic stress. In order to compute the literal listener's interpretation of utterances, we will start by computing this listener's inferences over which utterance the speaker intended. Suppose first that the listener perceives utterance a, with prosodic stress $s = 0$ (i.e. without prosodic stress). Then the distribution over intended utterances $P(u_i|u_p = a, s = 0)$ can be computed as follows, using Equation 3.2:

$$
\begin{aligned}
P(u_i = a|u_p = a, s = 0) &= \frac{P(u_i = a)P_N(a|a,0)}{\sum_{u_j} P(u_i = u_j)P_N(a|u_j,0)} \\
&= \frac{P(u_i = a)P_N(a|a,0)}{P(u_i = a)P_N(a|a,0) + P(u_i = b)P_N(a|b,0) + P(u_i = a \wedge b)P_N(a|a \wedge b,0)} \\
&= \frac{\frac{1}{3} \cdot 0.99}{\frac{1}{3} \cdot 0.99 + \frac{1}{3} \cdot 0.01 + \frac{1}{3} \cdot 0} \\
&= 0.99
\end{aligned}
$$

		u_i		
		a	b	$a \wedge b$
u_p	a	0.99	0.01	0
	b	0.01	0.99	0
	$a \wedge b$	0	0	1

Table B.1: Distribution over intended utterances u_i, given each perceived utterance u_p. The speaker is assumed to have not used prosodic stress.

		u_i		
		a	b	$a \wedge b$
u_p	a	0.995	0.005	0
	b	0.005	0.995	0
	$a \wedge b$	0	0	1

Table B.2: Distribution over intended utterances, when the speaker has used prosodic stress.

Given that the listener perceived utterance a without prosodic stress, the probability that the speaker intended utterance a is 0.99. In other words, the literal listener believes that they perceived the speaker's utterance accurately with probability 0.99. The probability that the speaker's intended utterance is b can be computed as follows:

$$\begin{aligned} P(u_i = b | u_p = a, s = 0) &= 1 - P(u_i = a | u_p = a, s = 0) \\ &= 0.01 \end{aligned}$$

When the literal listener perceives utterance b, the probabilities can be computed in an identical manner, due to the symmetry of the noise channel. In addition, because the utterance $a \wedge b$ is never corrupted under the noise channel (by assumption), if the listener perceives utterance $a \wedge b$, they always believe that they have accurately perceived the speaker. Table B.1 shows the listener's inferences about intended utterances, given each perceived utterance.

We next provide the literal listener's inferences in the case that the speaker has used prosodic stress. We assume that the use of prosodic stress decreases the noise rate by a factor of 2. When the listener perceives the utterance a, the probability that the speaker intended this utterance can be computed as follows:

$$P(u_i = a|u_p = a, s = 1) = \frac{P(u_i = a)P_N(a|a,1)}{\sum_{u_j} P(u_i = u_j)P_N(a|u_j,1)}$$
$$= \frac{P(u_i = a)P_N(a|a,1)}{P(u_i = a)P_N(a|a,1) + P(u_i = b)P_N(a|b,1) + P(u_i = a \wedge b)P_N(a|a \wedge b,1)}$$
$$= \frac{\frac{1}{3} \cdot 0.995}{\frac{1}{3} \cdot 0.995 + \frac{1}{3} \cdot 0.005 + \frac{1}{3} \cdot 0}$$
$$= 0.995$$

In other words, when the speaker uses prosodic stress the listener now believes that they have accurately perceived the speaker's utterance with probability 0.995. The remaining probabilities in this case are shown in Table B.2.

B.1.2 Literal interpretation of utterances

We have shown how the literal listener draws inferences about the speaker's intended utterance, in the presence of the noisy channel. This allows us to consider how the literal listener will *interpret* each utterance, i.e. what distribution over worlds the listener will assign to each utterance. We will first show how to compute $L_0(\{Alice\}\,|a,0)$, i.e. the probability that the listener assigns to world $\{Alice\}$ given the perceived utterance a and no prosodic stress. In order to use Equation 3.5, we first need to compute the quantity $K(w|a,0)$ for each world w. This is the probability that world w is literally consistent with the speaker's intended utterance. Using Equation 3.4 and the probabilities from Table B.1, the quantity $K(\{Alice\}\,|a,0)$ can be found as follows:

$$K(\{Alice\}\,|a,0) = \sum_{u_i} P(u_i|u_p = a, s = 0)\mathbb{1}_{\{Alice\} \in [\![u_i]\!]}$$
$$= P(u_i = a|u_p = a, s = 0)\mathbb{1}_{\{Alice\} \in [\![a]\!]} + P(u_i = b|u_p = a, s = 0)\mathbb{1}_{\{Alice\} \in [\![b]\!]}$$
$$= 0.99 \cdot 1 + 0.01 \cdot 0$$
$$= 0.99$$

		World w		
		$\{Alice\}$	$\{Bob\}$	$\{Alice, Bob\}$
u_p	a	0.495	0.005	0.5
	b	0.005	0.495	0.5
	$a \wedge b$	0	0	1

Table B.3: L_0 distribution over worlds, when the speaker has not used prosodic stress.

		World w		
		$\{Alice\}$	$\{Bob\}$	$\{Alice, Bob\}$
u_p	a	0.4975	0.0025	0.5
	b	0.0025	0.4975	0.5
	$a \wedge b$	0	0	1

Table B.4: L_0 distribution over worlds, when the speaker has used prosodic stress.

This means that when the listener hears utterance a without prosodic stress, they infer that, with probability 0.99, the literal meaning of the speaker's intended utterance is compatible with world $\{Alice\}$. The probabilities $K(w|a,0)$ for the other worlds w can be computed in a similar manner. Combining these probabilities with Equation 3.5, we can compute $L_0(\{Alice\}\,|a,0)$:

$$L_0(\{Alice\}\,|a,0) = \frac{P(\{Alice\})K(\{Alice\}\,|a,0)}{\sum_w P(w)K(w|a,0)}$$

$$= \frac{P(\{Alice\})K(\{Alice\}\,|a,0)}{P(\{Alice\})K(\{Alice\}\,|a,0) + P(\{Bob\})K(\{Bob\}\,|a,0) + P(\{Alice,Bob\})K(\{Alice,Bob\}\,|a,0)}$$

$$= \frac{\frac{1}{3}\cdot 0.99}{\frac{1}{3}\cdot 0.99 + \frac{1}{3}\cdot 0.01 + \frac{1}{3}\cdot 1}$$

$$= 0.495$$

The L_0 probabilities for other worlds w can be computed similarly, and are shown in Tables B.3 and B.4. Table B.3 shows the listener's interpretation when the speaker has not used prosodic stress, while Table B.4 shows the listener's interpretation when the speaker *has* used prosodic stress. When the speaker uses prosodic stress, the listener L_0's interpretation of their perceived utterance is closer to that utterance's literal meaning, than in the case that the speaker leaves the utterance unstressed. For example, when the listener hears utterance a used with stress, they only assign probability 0.0025 to the world $\{Bob\}$ (the world in which only Bob went to the store). This world, which is literally incompatible with utterance a, is assigned twice this probability (0.005) when the speaker does not use prosodic stress. When the listener hears the utterance a with stress, they infer that the speaker is more likely to have intended this utterance, and less likely to infer that the utterance b was actually intended. Thus, by placing stress on the utterance, the speaker can make the listener L_0 more likely to infer a world which is literally compatible with this utterance.

B.1.3 Speaker information utility

The speaker S_1 knows the state of the world w, and wants to communicate the answer to the QUD q to the listener L_0. First consider the speaker who knows the world $\{Alice\}$, i.e. that only Alice went to the store. There are two possible QUDs for this speaker: q_A (*Did Alice go to the store?*) and q_L (*Exactly who went to the store?*). Suppose that this speaker has QUD q_A. In order to compute the speaker probabilities according to Equation 3.10, we first need to compute the information utility of each utterance, using Equation 3.7. The information utility of the utterance a with prosodic stress 0 is given by:

$$\begin{aligned}
I_1(a,0|\{Alice\},q_A) &= -\log\frac{1}{R_1(a,0|\{Alice\},q_A)}\\
&= \log R_1(a,0|\{Alice\},q_A)\\
&= \log\sum_{w}\mathbb{1}_{q_A(w)=q_A(\{Alice\})}L_0(w'|a,0)\\
&= \log(\mathbb{1}_{q_A(\{Alice\})=\top}L_0(\{Alice\}|a,0)+\mathbb{1}_{q_A(\{Bob\})=\top}L_0(\{Bob\}|a,0)\\
&+\mathbb{1}_{q_A(\{Alice,Bob\})=\top}L_0(\{Alice,Bob\}|a,0))\\
&= \log(1\cdot 0.495+0\cdot 0.005+1\cdot 0.5)\\
&= \log(0.995)
\end{aligned}\tag{B.1}$$

We used the probabilities from Table B.3 to fill in the L_0 terms in this equation. There is one subtle part of this computation: the calculation of $R_1(a,0|\{Alice\},q_A)$, the probability that the listener L_0 will assign to the correct answer to the speaker's QUD. In order to

		Prosodic stress	
		0	1
u_p	a	$\log(0.995)$	$\log(0.9975)$
	b	$\log(0.505)$	$\log(0.5025)$
	$a \wedge b$	$\log(1)$	$\log(1)$

Table B.5: Each cell shows the information utility $I_1(u_p, s|\{Alice\}, q_A)$, i.e. the information utility of the perceived utterance and prosodic stress, for the speaker who knows that the true world is $\{Alice\}$ and wants to communicate QUD q_A.

calculate this probability, we sum over all of the worlds in which the speaker's QUD has the same answer as in the actual world. The actual world in this case is $\{Alice\}$, and the speaker's QUD is whether Alice went to the store, so the two worlds with the same answer to this question are $\{Alice\}$ and $\{Alice, Bob\}$ The computation shows that if the listener L_0 perceives utterance a with prosodic stress 0, then they will infer the correct answer the speaker's question with probability 0.995, and therefore the speaker will receive information utility equal to $\log(0.995)$. The information utilities for the other utterances are shown in Table B.5.

Given the information utilities for each individual utterance, we can compute the expected information utility from choosing utterance a with prosodic stress $s = 0$, using Equation 3.8:

$$\begin{aligned}
&\mathbb{E}_{P_N(\cdot|a,0)} I_1(\cdot|\{Alice\}, q_A) = \sum_{u_p} P_N(u_p|a,0) I_1(u_p|\{Alice\}, q_A) \\
&= P_N(a|a,0) I_1(a|\{Alice\}, q_A) + P_N(b|a,0) I_1(b|\{Alice\}, q_A) + P_N(a \wedge b|a,0) I_1(a \wedge b|\{Alice\}, q_A) \\
&= 0.99 \cdot \log(0.995) + 0.01 \cdot \log(0.505) + 0 \cdot \log(1) \\
&= 0.99 \cdot \log(0.995) + 0.01 \cdot \log(0.505)
\end{aligned} \tag{B.2}$$

Next, we will compute the speaker's expected information utility from choosing utterance a with prosodic stress $s = 1$. This will allow us to evaluate the relative value of using prosodic stress for this speaker. Again, we assume that the speaker is in world $\{Alice\}$ and has QUD q_A. The information utilities for each perceived utterance with prosodic stress

$s = 1$ are shown in Table B.5. Substituting the information utilities from this table into the computations in B.2, we can find the expected information utility:

$$\mathbb{E}_{P_N(\cdot|a,1)} I_1(\cdot|\{Alice\}, q_A) = 0.995 \cdot \log(0.9975) + 0.005 \cdot \log(0.5025) \tag{B.3}$$

Taking the difference between the two expected information utilities, we see

$$\mathbb{E}_{P_N(\cdot|a,1)} I_1(\cdot|\{Alice\}, q_A) - \mathbb{E}_{P_N(\cdot|a,0)} I_1(\cdot|\{Alice\}, q_A) \approx 0.01 \tag{B.4}$$

That is, the expected information utility from using prosodic stress is higher than the expected information utility from not using stress. Prosodic stress reduces the noise rate, and makes the listener more likely to accurately perceive the utterance a. As a result, this listener will be more likely to believe that Alice went to the store, and hence will be more likely to believe the correct answer to the speaker's QUD.

Equation B.4 shows the gain in information utility from using prosodic stress, for the speaker with QUD q_A. We will compare this to the gain in information utility for the speaker with QUD q_L. The speaker with QUD q_L wants to answer the question *Exactly who went to the store?* Suppose that the speaker knows that the true world is $\{Alice\}$. The information utility of the utterance a with prosodic stress 0 can be calculated by:

$$\begin{aligned}
& I_1(a,0|\{Alice\}, q_L) = \log R_1(a,0|\{Alice\}, q_L) \\
& = \log \sum_{w} \mathbb{1}_{q_L(w)=q_L(\{Alice\})} L_0(w'|a,0) \\
& = \log(\mathbb{1}_{q_L(\{Alice\})=\{Alice\})} L_0(\{Alice\}|a,0) + \mathbb{1}_{q_L(\{Bob\})=\{Alice\}} L_0(\{Bob\}|a,0) \\
& + \mathbb{1}_{q_L(\{Alice,Bob\})=\{Alice\}} L_0(\{Alice,Bob\}|a,0)) \\
& = \log(1 \cdot 0.495 + 0 \cdot 0.005 + 0 \cdot 0.5) \\
& = \log(0.495)
\end{aligned} \tag{B.5}$$

As in the calculation in B.1, we use the probabilities from Table B.3 to fill in the L_0 terms in this equation. The crucial difference between this calculation and the one in B.1 is in how the term $R_1(a,0|\{Alice\}, q_L)$ is computed. This is the probability that the listener L_0

		Prosodic stress	
		0	1
u_p	a	log(0.495)	log(0.4975)
	b	log(0.005)	log(0.0025)
	$a \wedge b$	log(0)	log(0)

Table B.6: Each cell shows the information utility $I_1(u_p, s|\{Alice\}, q_L)$, i.e. the information utility of the perceived utterance and prosodic stress, for the speaker who knows that the true world is $\{Alice\}$ and wants to communicate QUD q_L.

assigns to the correct answer to the QUD q_L. The correct answer in this case is that Alice, and only Alice, went to the store. In order to compute $R_1(a, 0|\{Alice\}, q_L)$, we therefore sum over worlds in which only Alice went to the store. There is only one world satisfying this condition: $\{Alice\}$. The probability that the listener assigns to the correct answer is therefore the probability that they assign to this world, 0.495. The remaining information utilities for this speaker are shown in Table B.6.

The expected information utility from not using prosodic stress can be found using the information utilities in Table B.6:

$$\mathbb{E}_{P_N(\cdot|a,0)} I_1(\cdot|\{Alice\}, q_L) = 0.99 \cdot \log(0.495) + 0.01 \cdot \log(0.005) \tag{B.6}$$

The speaker's information utilities from using prosodic stress are also shown in Table B.6. Using these utilities, we can compute the expected information utility in this case:

$$\mathbb{E}_{P_N(\cdot|a,1)} I_1(\cdot|\{Alice\}, q_L) = 0.995 \cdot \log(0.4975) + 0.005 \cdot \log(0.0025) \tag{B.7}$$

The difference between the expected information utilities can be found by:

$$\mathbb{E}_{P_N(\cdot|a,1)} I_1(\cdot|\{Alice\}, q_L) - \mathbb{E}_{P_N(\cdot|a,0)} I_1(\cdot|\{Alice\}, q_L) \tag{B.8}$$

$$= 0.995 \cdot \log(0.4975) + 0.005 \cdot \log(0.0025) - 0.99 \cdot \log(0.495) - 0.01 \cdot \log(0.005) \tag{B.9}$$

$$\approx 0.005 \cdot \log(0.0025) - 0.01 \cdot \log(0.005) \tag{B.10}$$

$$\approx 0.02 \tag{B.11}$$

In other words, for the speaker who has QUD q_L, using prosodic stress leads to an increase of 0.02 in expected information utility. This increase is approximately twice what was found for the speaker with QUD q_A, in Equation B.4, leading us to conclude:

$$\mathbb{E}_{P_N(\cdot|a,1)} l_1(\cdot|\{Alice\}, q_L) - \mathbb{E}_{P_N(\cdot|a,0)} l_1(\cdot|\{Alice\}, q_L) > \mathbb{E}_{P_N(\cdot|a,1)} l_1(\cdot|\{Alice\}, q_A) - \mathbb{E}_{P_N(\cdot|a,0)} l_1(\cdot|\{Alice\}, q_A) \tag{B.12}$$

Using prosodic stress therefore leads to a greater increase in information value for the speaker with QUD q_L than for the speaker with QUD q_A.

B.1.4 Speaker utterance choice

We have so far shown how to compute the expected information utilities for the speaker in world $\{Alice\}$ with either the polar-QUD or the list-QUD. We will now compute the distribution over utterances and prosodies for this speaker. Suppose that the speaker wants to communicate the answer to the polar-QUD q_A. Using Equations 3.3, B.2, and B.3, we can compute the overall utilities for this speaker:

$$\begin{aligned} U_1(a,0|\{Alice\}, q_A) &= \mathbb{E}_{P_N(\cdot|a,0)} l_1(\cdot|\{Alice\}, q_A) - c(a) - c(0) && \text{(B.13)} \\ &= 0.99 \cdot \log(0.995) + 0.01 \cdot \log(0.505) - c(a) - c(0) && \text{(B.14)} \end{aligned}$$

$$U_1(a,1|\{Alice\}, q_A) = 0.995 \cdot \log(0.9975) + 0.005 \cdot \log(0.5025) - c(a) - c(1) \tag{B.15}$$

Similarly, we can compute the utilities for the speaker with the list-QUD q_L with Equations B.6 and B.7:

$$U_1(a,0|\{Alice\}, q_L) = 0.99 \cdot \log(0.495) + 0.01 \cdot \log(0.005) - c(a) - c(0) \tag{B.16}$$

$$U_1(a,1|\{Alice\}, q_L) = 0.995 \cdot \log(0.4975) + 0.005 \cdot \log(0.0025) - c(a) - c(1) \tag{B.17}$$

Given the overall utilities from each utterance-prosody pair, we can find the distribution over the speaker S_1's choices. Following Equation 3.10, the probability that the speaker with QUD q_A will use utterance a with prosodic stress is given by:

$$S_1(a,1|\{Alice\},q_A) = \frac{e^{\lambda U_1(a,1|\{Alice\},q_A)}}{\sum_{u_i,s} e^{\lambda U_1(u_i,s|\{Alice\},q_A)}} \tag{B.18}$$

The *odds* of using prosodic stress given QUD q_A are therefore:

$$\frac{S_1(a,1|\{Alice\},q_A)}{1-S_1(a,1|\{Alice\},q_A)} < \frac{S_1(a,1|\{Alice\},q_A)}{S_1(a,0|\{Alice\},q_A)} \tag{B.19}$$

$$= e^{\lambda(U_1(a,1|\{Alice\},q_A)-U_1(a,0|\{Alice\},q_A))} \tag{B.20}$$

$$= e^{\lambda(\mathbb{E}_{P_N(\cdot|a,1)}I_1(\cdot|\{Alice\},q_A)-\mathbb{E}_{P_N(\cdot|a,0)}I_1(\cdot|\{Alice\},q_A)-c(1)+c(0))} \tag{B.21}$$

The inequality in Equation B.19 holds because $(1-S_1(a,1|\{Alice\},q_A)) > S_1(a,0|\{Alice\},q_A)$. This is a measure of how likely the speaker is to use prosodic stress given QUD q_A; higher odds indicate a higher likelihood.

The probability that the speaker with list-QUD q_L will use utterance a with prosodic stress is:

$$S_1(a,1|\{Alice\},q_L) = \frac{e^{\lambda U_1(a,1|\{Alice\},q_L)}}{\sum_{u_i,s} e^{\lambda U_1(u_i,s|\{Alice\},q_L)}} \tag{B.22}$$

$$\approx \frac{e^{\lambda U_1(a,1|\{Alice\},q_L)}}{e^{\lambda U_1(a,0|\{Alice\},q_L)} + e^{\lambda U_1(a,1|\{Alice\},q_L)}} \tag{B.23}$$

$$\tag{B.24}$$

The approximation in Equation B.23 holds because the speaker with QUD q_L receives very low utility from any utterances other than a (because these other utterances will communicate false answers to the QUD), and therefore these utterances contribute only negligibly to the sum in the denominator. This approximation implies that $1-S_1(a,1|\{Alice\},q_L) \approx S_1(a,0|\{Alice\},q_L)$, i.e. if the speaker does not use utterance a with prosodic stress, then they will use utterance a *without* stress (and not some other utterance). The odds of the speaker with QUD q_L using prosodic stress are:

$$\frac{S_1(a,1|\{Alice\},q_L)}{1-S_1(a,1|\{Alice\},q_L)} \approx \frac{S_1(a,1|\{Alice\},q_L)}{S_1(a,0|\{Alice\},q_L)} \tag{B.25}$$

$$= e^{\lambda(U_1(a,1|\{Alice\},q_L)-U_1(a,0|\{Alice\},q_L))} \tag{B.26}$$

$$= e^{\lambda(\mathbb{E}_{P_N(\cdot|a,1)}I_1(\cdot|\{Alice\},q_L)-\mathbb{E}_{P_N(\cdot|a,0)}I_1(\cdot|\{Alice\},q_L)-c(1)+c(0))} \tag{B.27}$$

In order to determine which QUD type will induce the speaker to use prosodic stress more often, we will compute the ratio of the two odds:

$$\frac{S_1(a,1|\{Alice\},q_L)/(1-S_1(a,1|\{Alice\},q_L))}{S_1(a,1|\{Alice\},q_A)/(1-S_1(a,1|\{Alice\},q_A))} > \frac{S_1(a,1|\{Alice\},q_L)/S_1(a,0|\{Alice\},q_L)}{S_1(a,1|\{Alice\},q_A)/S_1(a,0|\{Alice\},q_A)} \tag{B.28}$$

$$= \frac{e^{\lambda(\mathbb{E}_{P_N(\cdot|a,1)}I_1(\cdot|\{Alice\},q_L)-\mathbb{E}_{P_N(\cdot|a,0)}I_1(\cdot|\{Alice\},q_L)-c(1)+c(0))}}{e^{\lambda(\mathbb{E}_{P_N(\cdot|a,1)}I_1(\cdot|\{Alice\},q_A)-\mathbb{E}_{P_N(\cdot|a,0)}I_1(\cdot|\{Alice\},q_A)-c(1)+c(0))}} \tag{B.29}$$

$$= e^{\lambda((\mathbb{E}_{P_N(\cdot|a,1)}I_1(\cdot|\{Alice\},q_L)-\mathbb{E}_{P_N(\cdot|a,0)}I_1(\cdot|\{Alice\},q_L))-(\mathbb{E}_{P_N(\cdot|a,1)}I_1(\cdot|\{Alice\},q_A)-\mathbb{E}_{P_N(\cdot|a,0)}I_1(\cdot|\{Alice\},q_A)))} \tag{B.30}$$

$$> e^0 = 1 \tag{B.31}$$

The first inequality follows from Equation B.19, while the second inequality follows from Equation B.12, which showed that gain in expected information utility from prosodic stress is greater for the speaker with QUD q_L than the speaker with QUD q_A. This *odds ratio* indicates that the odds of using prosodic stress are greater with q_L than with q_A, and straightforwardly implies that the probability of using prosodic stress given q_L is greater than given q_A, i.e. $S_1(a,1|\{Alice\},q_L) > S_1(a,1|\{Alice\},q_A)$. The derivation shows that there is one factor driving this difference: the gain in information utility from using prosodic stress is greater with q_L than with q_A.

Bibliography

[1] Jay David Atlas. Topic/comment, presupposition, logical form and focus stress implicatures: The case of focal particles only and also. *Journal of Semantics*, 8(1-2):127–147, 1991.

[2] Matthew Aylett and Alice Turk. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and speech*, 47(1):31–56, 2004.

[3] Alan Clinton Bale. Scales and comparison classes. *Natural Language Semantics*, 19(2):169–190, 2011.

[4] J Barwise and J Perry. Situation and attitudes. 1983.

[5] David I Beaver and Brady Z Clark. *Sense and sensitivity: How focus determines meaning*, volume 12. John Wiley & Sons, 2009.

[6] Anton Benz, Gerhard Jäger, and Robert Van Rooij, editors. *Game theory and pragmatics*. Palgrave Macmillan, 2005.

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

[8] Leon Bergen, Roger Levy, and Edward Gibson. Verb omission errors: Evidence of rational processing of noisy language inputs. 2012.

[9] Leon Bergen, Roger Levy, and Noah D Goodman. Pragmatic reasoning through semantic inference, 2015.

[10] Andrea Bonomi and Paolo Casalegno. Only: Association with focus in event semantics. *Natural Language Semantics*, 2(1):1–45, 1993.

[11] Mara Breen, Evelina Fedorenko, Michael Wagner, and Edward Gibson. Acoustic correlates of information structure. *Language and Cognitive Processes*, 25(7-9):1044–1098, 2010.

[12] Daniel Büring. Topic. In Peter Bosch and Rob van der Sandt, editors, *Focus — Linguistic, Cognitive, and Computational Perspectives*, pages 142–165. Cambridge University Press, Cambridge, 1999.

[13] Colin F Camerer, Teck-Hua Ho, and Juin-Kuan Chong. A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, pages 861–898, 2004.

[14] Robyn Carston. Implicature, explicature, and truth-theoretic semantics. In Ruth Kempson, editor, *Mental Representations: The Interface Between Language and Reality*, pages 155–181. Cambridge University Press, 1988.

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

[16] Gennaro Chierchia, Danny Fox, and Benjamin Spector. Scalar implicature as a grammatical phenomenon. In Klaus von Heusinger, Claudia Maienborn, and Paul Portner, editors, *Semantics: An International Handbook of Natural Language Meaning*, volume 3, chapter 87. Berlin: Mouton de Gruyter, 2012.

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

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

[19] Meghan Clayards, Michael K Tanenhaus, Richard N Aslin, and Robert A Jacobs. Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, 108(3):804–809, 2008.

[20] L Jonathan Cohen. Some remarks on grice?s views about the logical particles of natural language. In *Pragmatics of natural languages*, pages 50–68. Springer, 1971.

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

[22] Judith Degen, Michael Franke, and Gerhard Jäger. Cost-based pragmatic inference about referential expressions. In Markus Knauff, Michael Pauen andNatalie Sebanz, and Ipke Wachsmuth, editors, *Proceedings of the Annual Meeting of the Cognitive Science Society*, pages 376–381, 2013.

[23] Laura C Dilley and Mark A Pitt. Altering context speech rate can cause words to appear or disappear. *Psychological Science*, 2010.

[24] Danny Fox. Free choice and the theory of scalar implicatures. In Uli Sauerland and Penka Stateva, editors, *Presupposition and implicature in compositional semantics*, pages 71–120. Basingstoke: Palgrave Macmillan, 2007.

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

[26] Danny Fox and Roni Katzir. On the characterization of alternatives. *Natural Language Semantics*, 19(1):87–107, 2011.

[27] Danny Fox and Benjamin Spector. Economy and embedded exhaustification. *Natural Language Semantics*, Forthcoming.

[28] Michael C Frank and Noah D Goodman. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998, 2012.

[29] Michael Franke. Signal to act: Game theory in pragmatics. 2009.

[30] Michael Franke and Judith Degen. Reasoning in reference games. *manuscript, Tübingen & Stanford*, 2015.

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

[32] Drew Fudenberg and Jean Tirole. Game theory. *Cambridge, MA*, 1991.

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

[34] Edward Gibson, Leon Bergen, and Steven T Piantadosi. Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, 110(20):8051–8056, 2013.

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

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

[37] H Paul Grice. Logic and conversation. *1975*, pages 41–58, 1975.

[38] Jeroen Groenendijk and Martin Stokhof. *Studies in the Semantics of Questions and the Pragmatics of Answers*. PhD thesis, University of Amsterdam, 1984.

[39] Carlos Gussenhoven. Focus, mode and the nucleus. *Journal of linguistics*, 19(02):377–417, 1983.

[40] Joy E Hanna, Michael K Tanenhaus, and John C Trueswell. The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language*, 49(1):43–61, 2003.

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

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

[43] Teck-Hua Ho, Colin Camerer, and Keith Weigelt. Iterated dominance and iterated best response in experimental" p-beauty contests". *American Economic Review*, pages 947–969, 1998.

[44] Laurence Horn. *A presuppositional analysis of only and even.* RI Binnick, 1969.

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

[46] Laurence Horn. *A Natural History of Negation.* University of Chicago Press, 1989.

[47] Laurence Horn. Exclusive company: Only and the dynamics of vertical inference. *Journal of semantics*, 13(1):1–40, 1996.

[48] Laurence Horn. Assertoric inertia and npi licensing. In *Chicago Linguistic Society*, volume 38, pages 55–82, 2002.

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

[50] Joachim Jacobs. Focus ambiguities. *Journal of Semantics*, 8(1-2):1–36, 1991.

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

[52] Justine T Kao, Leon Bergen, and Noah D Goodman. Formalizing the pragmatics of metaphor understanding. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, pages 719–724, 2014.

[53] Justine T Kao, Jean Y Wu, Leon Bergen, and Noah D Goodman. Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, 111(33):12002–12007, 2014.

[54] Christopher Kennedy. *Projecting the adjective: The syntax and semantics of gradability and comparison.* Routledge, 1999.

[55] Christopher Kennedy. Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and philosophy*, 30(1):1–45, 2007.

[56] Boaz Keysar, Shuhong Lin, and Dale J Barr. Limits on theory of mind use in adults. *Cognition*, 89(1):25–41, 2003.

[57] Ewan Klein. A semantics for positive and comparative adjectives. *Linguistics and philosophy*, 4(1):1–45, 1980.

[58] Manfred Krifka. *A compositional semantics for multiple focus constructions.* Springer, 1992.

[59] S-Y Kuroda. Indexed predicate calculus. *Journal of Semantics*, 1(1):43–59, 1982.

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

[61] Daniel Lassiter and Noah D Goodman. Adjectival vagueness in a bayesian model of interpretation. *Synthese*, 2015.

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

[63] Roger Levy. A noisy-channel model of rational human sentence comprehension under uncertain input. In *Proceedings of the conference on empirical methods in natural language processing*, pages 234–243. Association for Computational Linguistics, 2008.

[64] Roger Levy, Klinton Bicknell, Tim Slattery, and Keith Rayner. Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences*, 106(50):21086–21090, 2009.

[65] David Lewis. *Convention: A Philosophical Study*. Number 80. Harvard University Press, 1969.

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

[67] David Lewis. Scorekeeping in a language game. *Journal of philosophical logic*, 8(1):339–359, 1979.

[68] David K Lewis. *On the plurality of worlds*, volume 322. Cambridge Univ Press, 1986.

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

[70] David Marr. *Vision: A computational approach*. Freeman & Co., San Francisco, 1982.

[71] Marie-Christine Meyer. *Ignorance and Grammar*. PhD thesis, Massachusetts Institute of Technology, 2013.

[72] Richard Montague. The proper treatment of quantification in ordinary English. In Jaakko Hintikka, Julius Matthew Emil Moravcisk, and Patrick Suppes, editors, *Approaches to Natural Language*, pages 221–242. D. Reidel, Dordrecht, 1973.

[73] Richard Montague. The proper treatment of quantification in ordinary english. In Moravcsik Julius Hintikka, Jaakko and Patrick Suppes, editors, *Approaches to Natural Language*, pages 221–242. Reidel, Dordrecht, 1973.

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

[75] Aparna S Nadig and Julie C Sedivy. Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychological Science*, 13(4):329–336, 2002.

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

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

[78] Christopher Potts. The expressive dimension. *Theoretical linguistics*, 33(2):165–198, 2007.

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

[80] Craige Roberts. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Working Papers in Linguistics-Ohio State University Department of Linguistics*, pages 91–136, 1996.

[81] Richard M Roberts and Roger J Kreuz. Why do people use figurative language? *Psychological Science*, 5(3):159–163, 1994.

[82] Michael S Rochemont. *Focus in generative grammar*, volume 4. John Benjamins Publishing, 1986.

[83] Mats Rooth. Association with focus. 1985.

[84] Mats Rooth. A theory of focus interpretation. *Natural language semantics*, 1(1):75–116, 1992.

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

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

[87] Benjamin Russell. *Probabilistic reasoning and the computation of scalar implicatures*. PhD thesis, Brown University, 2012.

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

[89] Elisabeth Selkirk. Sentence prosody: Intonation, stress, and phrasing. In John Goldsmith, editor, *The Handbook of Phonological Theory*, pages 550–569. Basil Blackwell, London, 1999.

[90] Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[91] Nathaniel J. Smith and Roger Levy. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319, 2013.

[92] Dan Sperber and Deidre Wilson. *Relevance: Communication and Cognition*, volume 142. Cambridge, MA: Harvard University Press, 1986.

[93] Dan Sperber and Deirdre Wilson. *Relevance: Communication and cognition.* Harvard University Press, 1986.

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

[95] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction.* Cambridge Univ Press, 1998.

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

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

[98] Kai Von Fintel. *Restrictions on quantifier domains.* PhD thesis, University of Massachusetts, 1994.

[99] Arnim Von Stechow. Focusing and backgrounding operators. *Discourse Particles: Pragmatics & Beyond, Amsterdam: John Benjamins*, pages 37–84, 1991.