OXFORD UNIVERSITY PRESS | Journal of Semantics

Embedded implicatures as pragmatic inferences under compositional lexical uncertainty

Journal:	*Journal of Semantics*
Manuscript ID:	JS-15-01-002.R3
Manuscript Type:	Article
Keywords:	conversational implicatures, scalar implicatures, embedded implicatures, Bayesian pragmatics, experimental pragmatics

Embedded implicatures as pragmatic inferences under compositional lexical uncertainty*

Christopher Potts, Daniel Lassiter, Roger Levy, and Michael C. Frank

August 29, 2015

Abstract

How do comprehenders reason about pragmatically ambiguous scalar terms like *some* in complex syntactic contexts? In many pragmatic theories of conversational implicature, local exhaustification of such terms ('only some') is predicted to be difficult or impossible if the result does not entail the literal meaning, whereas grammatical accounts predict such construals to be robustly available. Recent experimental evidence supports the salience of these local enrichments, but the grammatical theories that have been argued to account for this evidence do not provide explicit mechanisms for weighting such construals against others. We propose a probabilistic model that combines previous work on pragmatic inference under 'lexical uncertainty' with a more detailed model of compositional semantics. We show that this model makes accurate predictions about new experimental data on embedded implicatures in both non-monotonic and downward-entailing semantic contexts. In addition, the model's predictions can be improved by the incorporation of neo-Gricean hypotheses about lexical alternatives. This work thus contributes to a synthesis of grammatical and probabilistic views on pragmatic inference.

1 Conversational implicature: Interacting with grammar

The linguistic forms that discourse participants exchange with each other routinely underrepresent the speaker's intended message and underdetermine the listener's inferences. Grice (1975) famously provided a philosophical framework for understanding the driving forces behind such pragmatic enrichment. At the heart of this framework are **conversational implicatures**: social, cognitively complex meanings that discourse participants create jointly in interaction.

Perhaps the best-studied examples of language users going beyond the literal semantics involve weak terms like *some* being strengthened to exclude their communicatively stronger alternatives, giving rise to construals like 'some and not all' or 'only some'. Such inferences are often called **scalar conversational implicatures** (SIs), and they are widely assumed to arise via the same social inferencing mechanisms that are at work in other implicatures. However, this assumption has always been controversial. Even Grice suggested that SIs might be closer to the grammar than other implicatures (p. 56; see also Levinson 2000; Sperber & Wilson 1995; Bach 2006), and

*All the data and code used in this paper are available at `https://github.com/cgpotts/pypragmods`

recent grammar-driven accounts are framed in direct opposition to an implicature analysis. For example, Chierchia et al. (2012: 2316) write, "the facts suggest that SIs are not pragmatic in nature but arise, instead, as a consequence of semantic or syntactic mechanisms". The ensuing debates have stimulated new insights, pushing researchers to identify and evaluate previously unnoticed consequences of the two broad positions.

Much of the debate between Gricean and grammar-driven accounts has centered around what we informally called **embedded implicatures** — cases where a pragmatically enriched interpretation seems to be incorporated into the compositional semantics. Such readings seem initially to demand implicature-enriched semantic representations. However, many of the relevant examples have received straightforward Gricean accounts in which semantic content and contextual assumptions interact to yield global implicatures that are meaning-equivalent to interpretations that would derive from local pragmatic enrichment (Russell 2006; Geurts 2009, 2011). This reduces the power of such examples to decide in favor of one side or the other.

Geurts & Pouscoulous (2009) and Chemla & Spector (2011) study weak scalar terms in a wide range of quantificational environments. They show that many of the attested listener inferences concerning such terms are amenable to Gricean treatments based on implicature calculation, with no need for such calculations to intrude on the semantics (see especially Geurts & Pouscoulous 2009: §8 and Chemla & Spector 2011: 361). However, they identify a class of examples that, if attested, would not admit of such a treatment: scalar terms in the scope of non-monotone quantifiers, as in *exactly one player hit some of his shots*. In such cases, exhaustification of the embedded quantifier (. . . *some but not all of his shots*) does not entail the literal meaning, whereas the Gricean implicature analysis of scalar terms can only strengthen literal meanings. Geurts & Pouscoulous's experiments fail to support enrichment in such contexts, whereas Chemla & Spector's suggest that it is possible. A number of recent papers have sought to make sense of these conflicting results (Clifton & Dube 2010; Geurts & van Tiel 2013; van Tiel 2014).

In this paper, we reproduce the central qualitative result of Chemla & Spector (2011) using more naturalistic experimental stimuli, a fully randomized between-subjects design to avoid unwanted inferences across critical items (Geurts & van Tiel 2013), and a more direct method of interpreting participants' responses. Like Chemla & Spector, we find that scalar terms in non-monotone environments support implicature inferences (though these seem not to be the preferred or most salient construals). In our view, this evidence points to an important role for compositional semantics in understanding implicatures.

To describe the complementary roles of grammar and pragmatics in embedded implicatures, we propose a model that both embraces the compositional insights of Chierchia et al. and characterizes how people arrive at such construals. This model is in the tradition of **rational speech act** models (Frank & Goodman 2012; Goodman & Stuhlmüller 2013) and **iterated best response** models (Franke 2009; Jäger 2012), and is a direct extension of the **compositional lexical uncertainty** model of Bergen et al. (2012) and Bergen et al. (2014). The model accounts for how discourse participants coordinate on the right logical forms (implicature-rich or not), seeking to retain the insights of Gricean accounts while paying close attention to the details of semantic composition.

We show that our model not only captures the qualitative pattern of implicature behaviors that Chemla & Spector found, but also makes quantitative predictions that are highly correlated with people's actual inferential behavior in context. In addition, we present evidence that these correlations can be improved if the set of refinements is lexically constrained, in keeping with broadly

neo-Gricean views of SIs (Horn 1972; Gazdar 1979a,b; Schulz & van Rooij 2006), though the precise nature of the true refinements remains a challenging open question. Our results suggest that the full theory of implicature depends substantively on the fine details of semantic composition *and* broader considerations of rational interaction. This is perhaps a departure from Grice's (1975) particular conception of pragmatic meaning, but it is well-aligned with his general theory of meaning and intention (Grice 1968, 1989; Grandy & Warner 2014). In view of our experimental results, the chief advantage of our model is that it makes quantitative predictions that are easily and rigorously linked with our human response patterns. In other words, the model makes predictions not only about which pragmatic inferences are possible but also about how likely those inferences are.

Our broader position is that grammar-driven accounts and Gricean accounts are not in opposition, but rather offer complementary insights. When communicating in natural languages, people are relying on linguistic conventions to try to identify and convey each other's intentions. All sides in the debate acknowledge this mix of grammatical and interactional factors. Grice's (1975) definition of conversational implicature is interactional, but his maxim of manner embraces a role for linguistic form. By introducing additional devices such as Horn scales, Neo-Griceans expand this role into areas Grice addressed with the maxims of quantity, quality, and relevance. Sperber & Wilson (1995) and Bach (1994) characterize many kinds of pragmatic enrichment as inferences about logical forms. And Chierchia et al. (2012) invoke pragmatic pressures to explain how speakers and listeners coordinate on whether to posit implicature-rich logical forms or more literal ones. Thus, there is substantially more consensus than the rhetoric often suggests.

2 Implicature, enrichment, and embedding

In this section, we describe embedded implicatures, seeking to identify the special theoretical challenges they pose. Under Grice's (1975) original definition, conversational implicature is an act of social cognition. The original definition is somewhat underspecified, and fleshing it out into a precise formulation is challenging (Hirschberg 1985), but the guiding idea seems clear. The listener assumes that the speaker is cooperative in the Gricean sense of rational interaction. However, the listener is confronted with an utterance U with content p that meets this assumption only if certain additional conditions are met. The listener can resolve this tension by positing that these conditions are in fact met; in many (but not all) cases, this means inferring that the speaker intended for the listener to infer the truth of a different but related proposition q. By this reasoning, the listener is able to reconcile the observation that the speaker chose to utter U with the assumption that the speaker is communicating cooperatively.

In the current work, we do not try to make the above description more rigorous. The model that we develop does not depend on an independently formulated definition of implicature, but rather seeks to derive such meanings from more basic considerations about how speakers and listeners reason about each other whenever they interact. Similarly, the model of Chierchia et al. (2012) is noncommittal about the reality of conversational implicatures per se. In that model, 'conversational implicature' can be seen as an informal label for a certain class of logical forms, rather than a conceptual primitive (see section 3 of this paper). With this in mind, we use the notion of conversational implicature only to articulate the central empirical focus of this paper — embedded scalar terms — and the associated challenges for formal pragmatic accounts.

On the classic Gricean account, SIs arise when the imperative 'Be as informative as is required'

(a subclause of the maxim of quantity) is in tension with another pragmatic pressure related to cooperative communication. The opposing force can take many forms, for example, relating to considerations of politeness, discretion, or secrecy, but it is usually attributed to the maxim of quality, which instructs speakers to say only what they have strong positive evidence for. For instance, imagine a sportscaster who has observed the outcome of a single round of a basketball tournament and is reporting on it as news. If the sportscaster says (1), then she will likely implicate that Player A did not make all of his shots.

(1) Player A hit some of his shots.

The SI follows from a straightforward application of the above ideas. We assume that the sportscaster is cooperative in the Gricean sense, and knowledgeable and forthcoming about the events. Why, then, did she opt for a weak statement like *Player A hit some of his shots* when a stronger statement like *Player A hit all of his shots* is available and would have been more informative? If knowledge is the only relevant consideration, it must be that she was prevented from using this stronger form because she does not know it to be true. Together with our assumption that she observed the full outcome, she can lack knowledge of this proposition only because it is false, leading to the implicated meaning that Player A did not hit all of his shots. In this way, a listener can enrich the speaker's message.

To make this example more concrete, suppose that we have two players, A and B, and that we care (for present purposes) only about whether each of them hit none, some but not all, or all of his shots. We can identify these (equivalence classes of) possible worlds with labels like `NA`, which means that Player A hit none of his shots and Player B hit all of his shots, and `SS`, which means that both players hit some but not all of their shots. There are $3^2 = 9$ such worlds. The literal semantics of (1) in this context is the proposition given in (2b). Our hypothesized implicature is (2c), the proposition that Player A did not hit all of his shots. The intersection of these two meanings delivers the communicated meaning, (2d).

(2)

a. Worlds:	`NN NS NA SN SS SA AN AS AA`	
b. Literal:	`SN SS SA AN AS AA`	'at least some'
c. Implicature:	`NN NS NA SN SS SA`	'not all'
d. Communicated:	`SN SS SA`	'only some'

There are many proposals for how to formalize this reasoning. The common theme running through all of them is that the implicature is accessible because it is an enrichment that strictly entails the original literal content — in this example, because the utterance's literal meaning and the implicature are combined into a stronger meaning by intersection. In Grice's terms, a general claim is further restricted by the interaction of quantity and quality.

The above reasoning extends to examples like (3), in which *some* is in the scope of a universal quantifier, though additional assumptions must be brought in to achieve a comparable implicature.

(3) Every player hit some of his shots.

Consider the potential enrichment of this sentence to convey that every player hit some but not all of his shots. This seems comparable to the construal we derived for (1), but it requires more assumptions. If we take the implicature to be the negation of the stronger alternative *every player hit all of his shots*, then the reasoning proceeds as in the first four lines of (4), which takes us to a

meaning (4d) that is consistent with one or the other of the players (but not both) having hit all of his shots. To arrive at the target meaning (every player hit some but not all of his shots), we must further assume an auxiliary premise beyond that required for (1). One example of such a premise is that of uniform outcomes (4e); there are many others that will do the job (Spector 2007b).

(4)	a. Worlds:	NN NS NA SN SS SA AN AS AA		
	b. Literal:	SS SA AS AA		‘all hit at least some’
	c. Implicature:	NN NS NA SN SS SA AN AS		‘not all hit all’
	d. Result:	SS SA AS		‘all hit some; not all hit all’
	e. Aux. premise:	NN SS AA		‘uniform outcomes’
	f. Communicated:	SS		‘all hit only some’

Though the need for an auxiliary premise is a noteworthy complication, it seems within the bounds of a Gricean account, and auxiliary premises like these might be independently justified (Russell 2006). As in the previous example, the communicated meaning is an enrichment of the literal content, and Gricean pressures and contextual assumptions deliver the stronger meaning. Geurts & Pouscoulous (2009) and Chemla & Spector (2011) home in on this common theme in scalar implicature calculation and use it to probe the scope and adequacy of the Gricean implicature framework. Examples like (5) are central to their discussions. This is a minimal variant of (3) with the subject universal determiner *every* replaced by *exactly one*.

(5) Exactly one player hit some of his shots.

Many people have the intuition that (5) can be used to describe a situation in which there is exactly one player who scored some but not all of his shots, which is consistent with some players having scored all of their shots. The reading is easy to characterize intuitively: one imagines that *some of his shots* has been locally enriched to *some but not all of his shots*, and that this enriched meaning is the semantic argument to the subject quantifier. What makes this reading notably different from, e.g., (3) is that it does not entail the literal reading, as we see in (6). The literal semantics is the proposition in (6b), whereas the content of the *. . . some but not all of his shots* (‘Local’) construal is (6c), which merely overlaps with it.

(6)	a. Worlds:	NN NS NA SN SS SA AN AS AA	
	b. Literal:	NS NA SN AN	‘exactly one hit at least some’
	c. Local:	NS SN SA AS	‘exactly one hit only some’

Any theory in which enriched scalar interpretations are always generated by intersection, as they are in classical Gricean and neo-Gricean accounts, will fail to arrive at (6c). Such theories head inexorably toward a refinement that excludes NA and AN, but they are essentially incapable of ‘introducing’ SA and AS. If such construals are possible, they must arise from other mechanisms.

The issue is even clearer when a scalar term is in the scope of a downward-monotone operator like *no*, as in *no player hit some of his shots*. In such cases, the embedded enrichment creates a meaning that is strictly entailed by (i.e., weaker than) the literal meaning:

(7)	a. Worlds:	NN NS NA SN SS SA AN AS AA	
	b. Literal:	NN	‘none hit some’
	c. Local:	NN NA AN AA	‘none hit only some’

Gricean theories predict that the 'local' enrichment of *some* to *only some* is unavailable as an implicature inference here, either because of the way pragmatic pressures interact or because *some* is held to be the strongest member of its scale in negative environments, leaving no room for further enrichment. Grammar-driven approaches have tended to agree with the basic empirical assumption, arguing that local enrichment is blocked in environments where it would strictly weaken the literal content (Chierchia 2006).

The empirical evidence is mixed but seems to support the accessibility of these local interpretations. Modifying an earlier design by Geurts & Pouscoulous (2009), Chemla & Spector used displays involving geometric patterns to assess whether interpreters could access local-enrichment readings of scalar terms in the scope of non-monotone and downward-monotone operators. Their findings suggest that local enrichment readings are available in both contexts, especially non-monotone ones. Skeptics of local enrichment have found grounds for challenging Chemla & Spector's findings (see section 5), but we believe that the theoretical challenges posed by embedded implicatures are real. In section 6, we describe a new experiment that reproduces the core qualitative findings of Chemla & Spector's studies.

3 CFS's grammar-driven model

This section briefly reviews the grammar-driven model of Chierchia et al. (2012) (henceforth CFS). The approach is inspired by those of Chierchia (2004), Spector (2007a), and Fox (2007, 2009). There are two central pieces to the account: a generally available function *ALT* that maps words and phrases to their alternatives, and a covert exhaustification operator *O*.

For *ALT*, the relevant notion of alternative is familiar from theories of questions and focus (Groenendijk & Stokhof 1984; Rooth 1985, 1992): we can assume, as a default, that the alternatives for an expression φ is some subset of the items in the same type-theoretic denotation domain as $[\![\varphi]\!]$, the meaning of φ. The precise value of the function *ALT* is context-dependent, and discourse participants are presumed to coordinate on it, just as they coordinate on the meanings of deictic or discourse-bound pronouns, elided phrases, and other pragmatically controlled free variables.

The effect of applying the basic exhaustification operator *O* to an expression φ in the context of a given *ALT* is shown in (8) (Spector 2007a; Fox 2007, 2009; Magri 2009; Chierchia et al. 2012).[1]

(8) $$O_{ALT}(\varphi) = [\![\varphi]\!] \sqcap \bigsqcap \{-q : q \in ALT(\varphi) \wedge [\![\varphi]\!] \not\sqsubseteq q\}$$

The *O* operator maps an expression φ to one that entails $[\![\varphi]\!]$ and excludes the denotations of expressions in $ALT(\varphi)$ that are not strictly weaker than $[\![\varphi]\!]$. When dealing with truth-functional expressions, we can regard $\sqcap$ as boolean conjunction and $\sqsubseteq$ as a material conditional, but the definition should be thought of as broad enough to include any kind of partial ordering (Hirschberg 1985: §4).

Part of the case for a grammar-driven view is that it uses pieces of semantic theory that are independently needed. In particular, exhaustification is at the heart of Groenendijk & Stokhof's (1984) theory of questions and their answers. The above operator is a common proposal for the meaning of *only* (for discussion: Rooth 1996; Büring & Hartmann 2001; Beaver & Clark 2008). Schulz & van Rooij

[1] This is not the operator that CFS ultimately favor, since it requires some implicit restrictions on allowable *ALT* functions in order to get the right inferences. The final version has the same form as (8) but further restricts *ALT*.

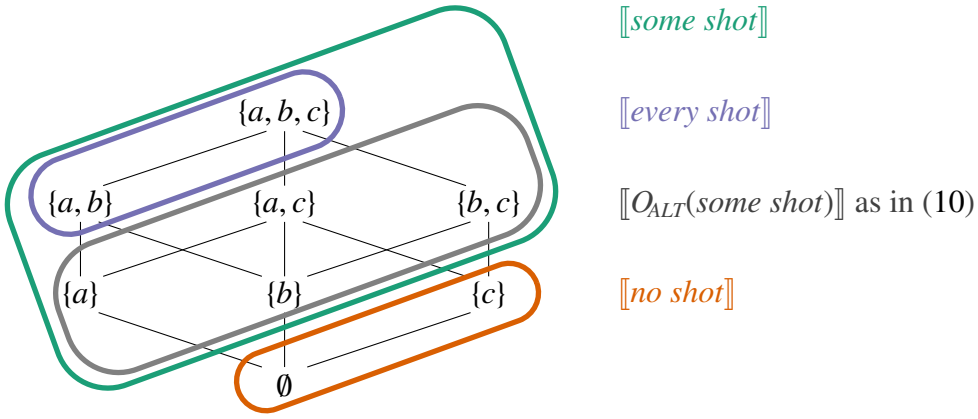

Figure 1: Given a domain $\{a, b, c\}$ with $[\![shot]\!] = \{a, b\}$, $[\![some\ shot]\!]$ is equal to the set of sets in the green box, $[\![every\ shot]\!]$ to the set of sets in the purple box, and $[\![no\ shot]\!]$ to the set of sets in the orange box. If *ALT*(*some shot*) contains $[\![every\ shot]\!]$, then *some shot* is refined to exclude the purple subset.

(2006) use exhaustification for implicature calculation (see also de Jager & van Rooij 2007). (For critical discussion, see Alonso-Ovalle 2008 and Gajewski 2012.) While CFS are cautious about making direction connections between *O* and these other phenomena (p. 2304), the correspondences are nonetheless noteworthy.

Those are the technical pieces. The proposal can then be summarized easily: *O* operators can optionally appear anywhere in the logical form of a sentence, perhaps subject to additional restrictions relating both to the comparative strength of the resulting logical form and to general pragmatic assumptions about the current conversational goals (see CFS: §4.6). To see the effects that this could have, let's return to the examples involving *some* that we reviewed in section 2. Simplifying slightly, let's suppose that *some shot* denotes the set of sets in (9) — the set of all sets *Y* that have a non-empty intersection with the set of shots.

(9) $[\![some\ shot]\!] = \{Y : [\![shot]\!] \cap Y \neq \emptyset\}$

Consider a domain of three entities $\{a, b, c\}$, and assume that $[\![shot]\!] = \{a, b\}$. Then the above is equivalent to the set of sets contained in the green box in figure 1. Now suppose that *ALT*(*some shot*) is defined as follows:

(10) $ALT(some\ shot) = \{[\![some\ shot]\!], [\![every\ shot]\!], [\![no\ shot]\!]\}$

a. $[\![some\ shot]\!]$ as in (9) (green circle in figure 1)

b. $[\![every\ shot]\!] = \{Y : [\![shot]\!] \subseteq Y\}$ (purple circle in figure 1)

c. $[\![no\ shot]\!] = \{Y : [\![shot]\!] \cap Y = \emptyset\}$ (orange circle in figure 1)

The presence of $[\![some\ shot]\!]$ has no effect because it is identical to the input. Similarly, all quantifiers that are weaker than the input have no effect if included in the *ALT* set. The presence of $[\![no\ shot]\!]$ has no effect because it contradicts the input, so its complement is weaker than the input. The presence of $[\![every\ shot]\!]$ will, though, be meaningful, as long as we assume that $[\![shot]\!] \neq \emptyset$. In that case, $O_{ALT}(some\ shot)$ will denote the subset in gray in figure 1. This is equivalent to the

intersection of $[\![\textit{some shot}]\!]$ and the complement of $[\![\textit{every shot}]\!]$ in the power set of the domain. In other words, it expresses *some and not all*, the intuitively implicature-rich interpretation.

Because O_{ALT} is embeddable, syntactic constituents like O_{ALT}(*some shot*) can appear in the scope of quantifiers. Implicature-rich versions of (1), (3), and (5) are thus available — potentially usable by speakers and inferable by listeners just like any other semantic resolution for an underspecified form in context.

As we noted in the introduction, CFS draw a firm rhetorical distinction between their proposal and the Gricean approach to pragmatics. They state, "the goal of this paper is to challenge the neo-Gricean approach to SIs" (p. 2303), and, as we said, they later write that "the facts suggest that SIs are not pragmatic in nature but arise, instead, as a consequence of semantic or syntactic mechanisms" (p. 2316). The sense in which their account reflects this position is clear: to characterize implicatures, we need not consider the interactional setting or try to model the speaker and hearer. Rather, we can just describe a specific class of logical forms.

This position is tempered by CFS's pervasive appeals to pragmatic reasoning, however. The authors' specific examples are generally placed in contexts that support the target implicatures by ensuring that they are relevant, informative, and truthful. They concede that "aspects of the Gricean picture are sound and effective" (p. 2299). And, in summarizing their account, they make explicit the role that pragmatics must play in helping discourse participants to coordinate on the right logical forms:

> one can capture the correlation with various contextual considerations, under the standard assumption (discussed in the very beginning of this paper) that such considerations enter into the choice between competing representations (those that contain the operator and those that do not). (p. 2317)

The coordination problem that Grice sought to solve therefore remains, in the following form. First, in CFS's theory, the discourse participants must coordinate on the nature of the function *ALT*. Second, because the language permits but does not require silent, embedded *O* operators in many positions, the speaker's signal frequently underdetermines her intended message; a given surface form *U* might be consistent with logical forms that encode implicatures and those that don't, depending on where *O* appeared. Crucially, the speaker must rely on the listener to select the right one. Overall, then, implicature calculation now amounts to reasoning about which logical form was intended. How this coordination happens has not been a focus of grammar-driven accounts, but the above quotation suggests that communicative pressures like those Grice identified guide the process.

Summarizing so far, we have evidence from Chemla & Spector's (2011) experiments that some implicatures require, in some sense, local enrichment of embedded content via enhanced logical forms. Traditional Gricean accounts seem unable to capture such cases, but such accounts excel at characterizing how speakers and listeners coordinate on implicatures in simpler cases. CFS, in contrast, define a model in which local calculation is immediate, but they do not venture an account of how discourse participants coordinate on the right logical forms when more than one is allowed by the grammar. Stepping back, we see that both the Gricean and grammar-driven accounts clearly have something to contribute. We now turn to the task of developing a synthesis of the two approaches: a model that formally implements pragmatic reasoning over complex, compositionally defined logical forms and that is able to achieve the readings that seem to demand local enrichment. The technical details of the compositional model are different from CFS's, and

the technical details of the pragmatic account are different from Grice, but we hope that it combines the best aspects of both approaches.

4 A compositional lexical uncertainty model

We now present our mixed semantic–pragmatic model, which can be seen as a conceptual fusion of the Montagovian semantic perspective in Lewis (1970), the signaling systems of Lewis (1969), the probabilistic rational speech acts perspective of Frank & Goodman (2012) and Goodman & Stuhlmüller (2013), the iterated best response model of Jäger (2007, 2012) and Franke (2009), and the Bayesian view of Gricean reasoning developed by Russell (2012). Our Python implementation of the model is available from the website for this paper.

The model we implement here is a direct extension of the compositional lexical uncertainty model of Bergen et al. (2012) and Bergen et al. (2014) (see also Lassiter & Goodman 2013, 2015, for a closely related variant). This model defines production and interpretation as recursive processes in which speakers and listeners reason jointly about the state of world and the precise interpretation of lexical items in context. Our extension simply allows for greater diversity in the semantic lexicon and includes more complex aspects of semantic composition. Thus, in many ways, our central theoretical result is that Bergen et al.'s model predicts embedded implicatures in non-monotone and downward-monotone contexts if it is combined with a full theory of semantic composition.

The model's crucial feature is **lexical uncertainty**. In semantics, we like to imagine that word meanings are fixed across speakers and contexts, but in fact they are often idiosyncratic and adaptable (Clark & Clark 1979; Clark 1997; Lascarides & Copestake 1998; Glucksberg 2001; for an overview and general discussion, see Wilson & Carston 2007). Thus, in our model, discourse participants are not presumed to share a single, fixed lexicon mapping word forms to meanings. Rather, they consider many such lexica, and their communicative behavior, in both production and interpretation, is guided by their best attempts to synthesize the information from these varied sources (Giles et al. 1991). Thus, in the sentences of interest, the discourse participants might entertain multiple senses for an embedded *some*, including not only its 'at least' meaning but also the 'only some' meaning that corresponds to its enrichment by scalar implicature. This uncertainty carries through the compositional semantics to deliver embedded implicature readings. From this perspective, Chierchia et al.'s model is conceptually very close to lexical uncertainty, in that it requires reasoning about the logical form that a speaker intends to convey; a given token of *some* can take on multiple senses depending on the presence and nature of silent embedded operators in the logical form. Our extension of Bergen et al.'s model shows how this uncertainty guides pragmatic reasoning, and it furthermore shows that the uncertainty need not be fully resolved in order for robust pragmatic inferences to go through.

4.1 Grammar fragment

Table 1 gives the intensional fragment that we use throughout the remainder of this paper, both to explain how our pragmatic model works and to conduct our experimental analyses in section 6. It is our base lexicon, subject to refinement as part of pragmatic inference.

Syntax	Denotation of the lefthand side of the syntax rule
$\text{N} \rightarrow \textit{person}$	$\{\langle w, x\rangle : x \text{ is a person in } w\}$
$\text{N} \rightarrow \textit{shot}$	$\{\langle w, x\rangle : x \text{ is a shot in } w\}$
$\text{V}_\text{T} \rightarrow \textit{hit}$	$\{\langle w, x, y\rangle : x \text{ hit } y \text{ in } w\}$
$\text{V}_\text{I} \rightarrow \textit{scored}$	$\{\langle w, x\rangle : \exists y\ x \text{ hit } y \text{ in } w\}$
$\text{V}_\text{I} \rightarrow \textit{cheered}$	$\{\langle w, x\rangle : x \text{ cheered in } w\}$
$\text{D} \rightarrow \textit{some}$	$\{\langle w, X, Y\rangle : \{x : \langle w, x\rangle \in X\} \cap \{y : \langle w, y\rangle \in Y\} \neq \emptyset\}$
$\text{D} \rightarrow \textit{every}$	$\{\langle w, X, Y\rangle : \{x : \langle w, x\rangle \in X\} \subseteq \{y : \langle w, y\rangle \in Y\}\}$
$\text{D} \rightarrow \textit{no}$	$\{\langle w, X, Y\rangle : \{x : \langle w, x\rangle \in X\} \cap \{y : \langle w, y\rangle \in Y\} = \emptyset\}$
$\text{D} \rightarrow \textit{exactly one}$	$\{\langle w, X, Y\rangle : \lvert\{x : \langle w, x\rangle \in X\} \cap \{y : \langle w, y\rangle \in Y\}\rvert = 1\}$
$\text{NP} \rightarrow \textit{Player A}$	$\{\langle w, Y\rangle : \mathtt{a} \in \{x : \langle w, x\rangle \in Y\}\}$
$\text{NP} \rightarrow \textit{Player B}$	$\{\langle w, Y\rangle : \mathtt{b} \in \{x : \langle w, x\rangle \in Y\}\}$
$\text{NP} \rightarrow \textit{Player C}$	$\{\langle w, Y\rangle : \mathtt{c} \in \{x : \langle w, x\rangle \in Y\}\}$
$\text{NP} \rightarrow \text{D N}$	$\{\langle w, Y\rangle : \langle w, [\![\text{N}]\!], Y\rangle \in [\![\text{D}]\!]\}$
$\text{VP} \rightarrow \text{V}_\text{T}\ \text{NP}$	$\{\langle w, x\rangle : \{\langle w, y\rangle : \langle w, x, y\rangle \in [\![\text{V}_\text{T}]\!]\} \in [\![\text{NP}]\!]\}$
$\text{VP} \rightarrow \text{V}_\text{I}$	$[\![\text{V}_\text{I}]\!]$
$\text{S} \rightarrow \text{NP VP}$	$\{w : \langle w, [\![\text{VP}]\!]\rangle \in [\![\text{NP}]\!]\}$

Table 1: Interpreted grammar fragment. The left column defines a context-free grammar, and the right column gives its recursive interpretation in an intensional model $\langle D, W, [\![\cdot]\!]\rangle$, where D is a set of entities, W is a set of possible worlds, and $[\![\cdot]\!]$ is a semantic interpretation function. Notational conventions: $x, y \in D$, $w \in W$, and $X, Y \subseteq (W \times D)$.

The formal presentation is influenced by that of Muskens (1995): all of the denotations are sets, and the rules of semantic composition (the final four lines) combine them using operations that are formally akin to functional application. Our motivation for this less familiar presentation is that it makes it easy to define a uniform notion of refinement throughout the lexicon.

4.2 Refinement

The grammar in table 1 contains both lexical entries and rules of semantic combination. We assume that the rules are fixed. The lexical entries, on the other hand, are merely a starting point for linguistic communication — a set of somewhat negotiable conventions. You might assume that *couch* and *sofa* are synonymous, but if I say "It's a couch but not a sofa", you'll learn something about my lexical representations and perhaps adjust your own accordingly for the purposes of our interaction. If a speaker uses the phrase *synagogues and other churches*, then the listener can conclude that the speaker regards a synagogue as a kind of church, via the presuppositional nature of the phrase. Conversely, if the speaker says *church or synagogue*, the listener receives a weak signal that the speaker regards those two terms as disjoint, via the pressure for disjuncts to be exclusive (Hurford 1974). Chemla (2013) and Potts & Levy (2015) explicitly investigate such

listener implicatures and how they can be anticipated and potentially forestalled by speakers.

The 'lexical uncertainty' aspects of our model are designed to capture this variability. The core notion is that of lexical **refinement**, as defined in (11) following Bergen et al. (2014):

(11) a. Let φ be a set-denoting expression. R is a **refinement** of φ iff $R \neq \emptyset$ and $R \subseteq [\![\varphi]\!]$.

b. $\mathcal{R}_c(\varphi)$, the set of refinements for φ in context c, is constrained so that $[\![\varphi]\!] \in \mathcal{R}_c(\varphi)$ and $\mathcal{R}_c(\varphi) \subseteq \wp([\![\varphi]\!]) - \emptyset$

The full possible refinement space for a lexical item is the power set of its denotation minus the empty set. In a functional presentation of the interpreted fragment, this could instead be defined in terms of the subfunctions of a given denotation using a cross-categorical notion of entailment. With (11b), we allow that contexts can vary in how much of the full refinement space they utilize. They can be as small as the original denotation (in which case the uncertainty is eliminated), or as large as the full power set (minus the empty set).

The guiding idea is that, in interaction, pragmatic agents reason about possible refinements of their lexical items, with the base lexical meaning serving as a kind of anchor to which each word's interpretation is loosely tethered. Intuitively, one can imagine that part of what it means to be a responsible interlocutor is to make inferences, based on the speaker's behavior, not only about the world information she would like to convey, but also about the precise meanings she intends the words she is using to carry in the context of the interaction.

As we noted above, CFS's model embodies a kind of semantic uncertainty very similar to that considered here. For any given expression that one hears, the speaker might have in mind its literal content $[\![\varphi]\!]$ or one of the many enrichments available with $O_{ALT}(\varphi)$ for different choices of *ALT*. Similarly, we admit the trivial refinement $R = [\![\varphi]\!]$ as well as enrichments (technically, subsets) of it. The major technical difference lies in how these sets of denotations enter into the compositional semantics. For CFS, the alternatives all contribute to a single denotation, whereas our model keeps the alternatives separate during semantic composition, synthesizing them only for pragmatic inference. In terms of figure 1, we saw that CFS's theory uses O_{ALT} to create a single refined meaning for *some shot*, represented by the set of sets in the gray box ('some, not all shots'). Our theory of refinement could create one lexicon for every non-empty subset of the green box. So, in addition to considering 'some, not all, shots', we admit lexica that produce $[\![\textit{some shot}]\!] = \{\{a, b, c\}\}$ ('every shot'), lexica that produce $[\![\textit{some shot}]\!] = \{\{a, b, c\}, \{a\}\}$ (no obvious paraphrase), and so forth. These are all potential results of $O_{ALT}(\textit{some shot})$ for some choice of *ALT*, and our theory can be regarded as one that reasons in terms of all of these options.

4.3 Pragmatic reasoning

Our pragmatic model combines the logical grammar of section 4.1 with the lexical refinements of section 4.2. The basic ingredients are given in (12). We take as given a context c, an interpreted fragment $\langle \mathcal{G}, D, W, [\![\cdot]\!] \rangle$ as in table 1, with context free grammar $\mathcal{G}$, a domain of entities D, a set of worlds W, an interpretation function $[\![\cdot]\!]$ interpreting expressions of $\mathcal{G}$ in these domains, and a refinement function $\mathcal{R}_c(\varphi)$ that is defined for all lexical items in $\mathcal{G}$. For convenience, we assume that W is finite; this simplifies the definition of the probability measures but is not otherwise crucial.

(12) a. M is a subset of the proposition-denoting expressions generated by $\mathcal{G}$. It is augmented with a null message $\mathbf{0}$ such that $[\![\mathbf{0}]\!] = W$.

b. $\mathbf{L} = \left\{\mathcal{L} : \text{ for all } w \in W, \mathcal{L}(\mathbf{0}, w) = 1, \text{ and for all } m \in M, \{w : \mathcal{L}(m, w) = 1\} \in \mathcal{R}_c(m)\right\}$

c. $P : \wp(W) \mapsto [0, 1]$ is a prior probability distribution over sets of worlds. (For notational convenience, we abbreviate $P(\{w\})$ as $P(w)$.)

d. $C : M \mapsto \mathbb{R}$ is a cost function on messages. For lexical items, costs are specified. For a nonterminal node A with daughters $B_1 \ldots B_n$, $C(A) = \sum_{i=1}^{n} C(B_i)$.

e. $P_{\mathbf{L}} : \wp(\mathbf{L}) \mapsto [0, 1]$ is a prior probability distribution over sets of lexica. (For notational convenience, we abbreviate $P_{\mathbf{L}}(\{\mathcal{L}\})$ as $P_{\mathbf{L}}(\mathcal{L})$.)

In this paper, we do not bias the prior distribution over states P or the prior distribution over lexica $P_{\mathbf{L}}$ in any way, assuming them to be flat. Since we do not have experimental measurements for the priors, this seems like the safest option. (For techniques for measuring and manipulating state priors, see Frank & Goodman 2012 and Stiller et al. 2011.) Similarly, we do not explore different cost functions on non-null messages, assuming all costs to be zero.[2] Our cost functions play a role only in disfavoring the 'null message' $\mathbf{0}$, which is stipulated to be true in all worlds in all lexica.

In the context of our model, the set of messages M creates a space of alternative utterances that can drive complex pragmatic reasoning, as we will see in section 4.4. However, while these alternatives play a crucial role in capturing implicatures, they do not suffice for embedded ones. Thus, our focus is on the space of lexica defined by (12b) given a certain set of relevant alternative messages, as in (12a). Clause (12b) specifies all of the possible lexica $\mathcal{L}$ given the original interpretation function $[\![\cdot]\!]$ and $\mathcal{R}_c$. It is the space opened up by these constructs that allows us to predict where and how embedded implicatures will be perceived as salient. It should be noted in this context that our decision to refine only lexical items, as in (12b), is made only for simplicity. We could also allow arbitrary words and phrases to be refined, as CFS in effect do.

With this background in place, we now specify the core lexical uncertainty model. It consists of three inter-related agents, as defined in (13). The agents are defined in terms of the cost function C, the state prior P, and the lexica in $\mathbf{L}$. We assume throughout that m is any message in M, w is any state in W, and $\mathcal{L}$ is any lexicon in the set $\mathbf{L}$.[3]

(13) a. $l_0(w \mid m, \mathcal{L}) \propto \mathcal{L}(m, w)P(w)$

b. $s_1(m \mid w, \mathcal{L}) \propto \exp\left(\log l_0(w \mid m, \mathcal{L}) - C(m)\right)$

c. $L(w \mid m) \propto P(w) \sum_{\mathcal{L} \in \mathbf{L}} P_{\mathbf{L}}(\mathcal{L}) s_1(m \mid w, \mathcal{L})$

The first two agents, l_0 and s_1, are fixed-lexicon agents, and the final listener L reasons over all of the lexica in $\mathbf{L}$. The most basic agent is l_0. It defines a conditional distribution over worlds w given messages m. It does this by simply combining the truth conditions, given numerically as $\mathcal{L}(m, w)$, with the state prior. Where $\mathcal{L}(m, w) = 1$, the value is proportional to the state prior value

[2] The model is mathematically invariant to across-the-board additive transformations of message costs, so assuming all non-null messages to have zero cost loses no generality.

[3] $P(a \mid b) \propto F(a)$ is read 'the value $P(a \mid b)$ is proportional to the value $F(a)$'. The exact value of $P(a \mid b)$ can always be obtained by dividing $F(a)$ by the normalizing constant $Z = \sum_{a'} F(a')$ so long as this sum is finite, which is guaranteed to be the case in the class of models defined in (12).

$P(w)$; where $\mathcal{L}(m, w) = 0$, the value is 0. So this is just the semantics turned into a probability distribution for the sake of decision making; the intuitive idea is that the agent hears m and estimates the relative likelihood of worlds on that basis.

The speaker agent s_1 is already a pragmatic agent, in the sense that it reasons not about the lexicon directly but rather about how the listener will reason about the lexicon. The speaker observes a state w and chooses messages on that basis. The logarithm and exponentiation in this definition allow us to include real-valued costs; where the costs are all 0, it reduces to $s_1(m \mid w) \propto l_0(w \mid m)$, by the identity $x = \exp(\log(x))$.[4]

Some comment is in order regarding the role of the null message in the model. Technically, $\mathbf{0}$ allows us to explore the full space of refinements for messages while guaranteeing that, for every possible speaker's observed state w, there is some compatible message m such that $\mathcal{L}(m, w) = 1$. Without this, the speaker distribution $s_1(m \mid w, \mathcal{L})$ would not be defined. There are a few alternative methods for addressing this technical issue. Bergen et al. (2012) admit only lexica in which the speaker has at least one true message for every state; Bergen et al. (2014) briefly consider giving false states tiny positive probability; and Jäger (2012) defines a belief-revision step to handle comparable situations in the context of the iterated best-response model. The null-message approach has qualitatively similar behavior to these other approaches, and we favor it here because it is technically simpler to implement.[5] We set $C(\mathbf{0}) = 5$ throughout the paper, but changing this value does not change our qualitative predictions. (See also Appendix A.)

Our pragmatic listener is defined in (13c). This agent resembles the literal listener l_0, but it sums over all of the inferences defined by the lexicon-specific agents s_1 and l_0. It additionally incorporates the state prior, as l_0 does, and the prior over lexica. This is the agent that we use to characterize listener inferences and define our predictions about our experimental findings.

We have presented the compositional lexical uncertainty model in its simplest form, but we have gone beyond Bergen et al. in three respects. We give a more complete treatment of semantic composition, we allow uncertainty in the denotations of lexical items of a wider range of semantic types, and we entertain the possibility of restrictions on the set of possible refinements. However, many other elaborations of models in this space are possible (Goodman & Lassiter 2015; Smith et al. 2013; Kao et al. 2014b; Potts & Levy 2015). Two particular elaborations are highly salient given the prior literature. First, one could allow further iteration beyond L, defining speaker and listener agents analogously to their fixed-lexicon counterparts. This can amplify existing pragmatic inferences and create new ones (Bergen et al. 2014; Vogel et al. 2014; Potts & Levy 2015). Second, one can include a real-valued temperature parameter λ in the speaker agents to control how greedily they try to extract information from the agent they are reasoning about, with higher λ values leading to more aggressive inferential strategies (Sutton & Barto 1998). This too can radically reshape the agents' behavior. In appendix A, we explore the consequences of these elaborations

[4] We could equivalently define an alternative cost function C' ranging over $[0, \infty)$ such that $C'(m) = e^{C(m)}$, and then replace (13b) with $s_1(m \mid w, \mathcal{L}) \propto l_0(w \mid m, \mathcal{L})C'(m)$.

[5] A closely related alternative, technically more complex but perhaps more pretheoretically transparent, would be to posit a collection of "null" messages, one for each speaker's observed state, each admitting only that state, and each having a considerably higher cost than all the non-null messages. This alternative has the interpretation that the null messages constitute the collection of more precise but much more prolix utterances the speaker might have used to describe her observation state. The behavior of this alternative approach would be qualitatively the same as ours: the specialization of each null message for a unique world state would strengthen its appeal for s_1, but its high cost would countervail that appeal.

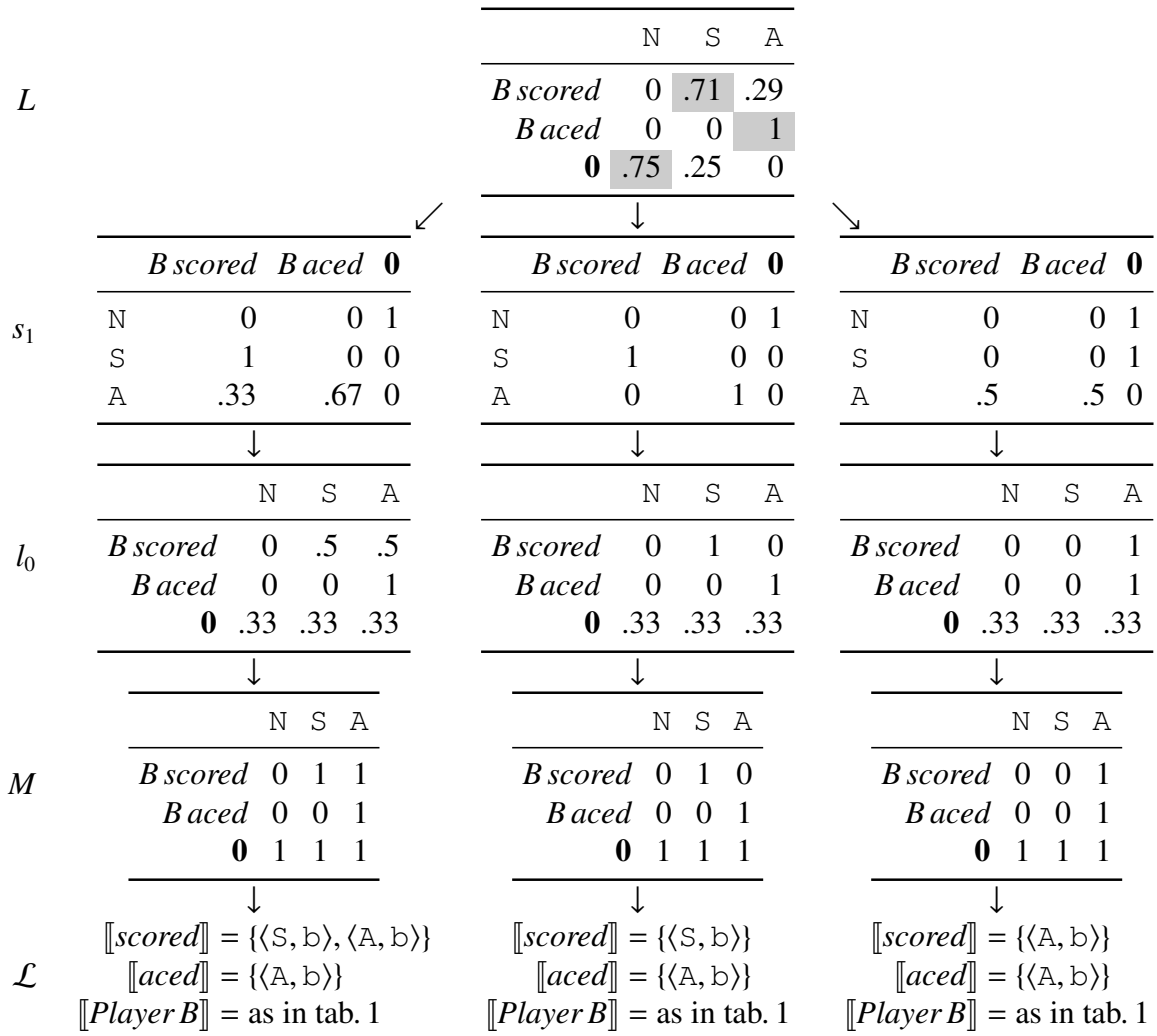

Figure 2: Simple scalar inference. We assume a flat prior over states and lexica. $C(\mathbf{0}) = 5$, and $C(m) = 0$ for the other messages. The uncertainty listener L infers that the general term *scored* excludes its specific counterpart *aced* in this context.

for modeling the pattern of data we observed.

4.4 Illustrations

Our first illustration, given in figure 2, is designed solely to reveal details about how the agents interact to produce enriched construals. (This first illustration is isomorphic to the example covered in section 4.4 of Bergen et al. 2014.) We assume that the domain consists of just one entity, b, and that the only intensional distinction of interest is whether b scored none of his shots (world N), some but not all of his shots (S), or all of his shots (A). The action is in the relationship between the two predicates *scored* and *aced*: we define $[\![\textit{scored}]\!] = \{\langle \mathtt{S}, \mathtt{b}\rangle, \langle \mathtt{A}, \mathtt{b}\rangle\}$ and $[\![\textit{aced}]\!] = \{\langle \mathtt{A}, \mathtt{b}\rangle\}$. Thus, *aced* strictly entails *scored*, creating the potential for an SI.

To keep the example compact, we let $\mathcal{R}_c(\textit{Player B}) = \{[\![\textit{Player B}]\!]\}$. Since *aced* already denotes a singleton set, it has no space for further refinement. However, *scored* has two further refinements. This gives rise to the three lexica in the bottom row of figure 2. Using the fixed rules of semantic composition, these lexica determine the messages *Player B scored* and *Player B aced*. The literal listener l_0 turns the denotations of these messages into conditional distributions over states given messages. The prior over states is flat in this example, so this calculation just evenly divides the probability mass over the true states. The pragmatic speaker responds to this agent. Finally, our uncertainty listener sums over these three speakers. This listener achieves an SI in the following nuanced, probabilistic sense (Russell 2012: §2). Hearing *Player B scored* leads this listener to assume that the most probable state is `S`. The probability is not 1, so uncertainty remains. However, if this listener is compelled to make a categorical decision about the intended meaning of the utterance, he will choose this enriched construal, and he will rightfully feel deceived if the world state turns out to be `A` or (worse) `N` instead. In this way, the model characterizes the uncertainty surrounding implicature inferences (Hirschberg 1985) and the ways in which this uncertainty relates to decision making.

Lexical uncertainty is not required to achieve this result. If we allow no meanings to be refined, then we deal with the singleton set of lexica containing only the leftmost lexicon. In this small space, the model shares deep affinities with the Bayesian model of Gricean reasoning given by Russell (2012); it is effectively equivalent to the rational speech act model of Frank & Goodman (2012) (potentially with small differences relating to how the prior over states is incorporated); and it can be seen as a more thoroughly probabilistic version of the iterated best response model (Franke 2009; Jäger 2007, 2012). Nonetheless, the example illuminates how the lexical uncertainty model works. As the downward arrows indicate, it is useful to start conceptually from L. This agent effectively reasons in Gricean terms about three separate lexica; the alternation from speaker to listener and down to the lexicon mirrors the nested belief structure of Grice's original definition of implicature (sketched at the start of section 2).

Even though we assume an even prior over lexica, useful biases emerge because the space of lexica is structured: there are no lexica in which *aced* is consistent with `S`, but there are two in which *scored* is. This bias carries through the computation to create a strong final bias for the implicature inference. For further discussion of this important point, we refer to Bergen et al. 2014, where it is shown to be essential to generating implicatures based on the principle that marked forms signal marked meanings and unmarked forms signal unmarked meanings (McCawley 1978; Horn 1984; Blutner 1998; Levinson 2000).

The lexical uncertainty aspects of the model are a rich source of implicatures, and they are the key to achieving local implicatures of the sort reviewed in section 2 above. However, as the fixed lexicon versions of the model make clear, the recursive nature of the agents suffices for many kinds of enrichment assuming the space of alternative messages M is chosen properly. Even with a single lexicon, we have a listener reasoning about a speaker reasoning about the literal interpretive semantics, which creates forces for removing semantic overlap among the alternative messages. One powerful illustration of this comes from Sauerland (2001, 2004), who studies the implicatures of sentences like *Player A hit some of his shots or cheered*, in which the weak scalar term *some of his shots* is nested inside the weak connective *or*. The guiding intuition is that the sentence is most prominently construed as entailing that Player A did not make all of his shots and that Player A did not both make shots and cheer. Sauerland's insight is that these entailments are within reach of traditional neo-Gricean reasoning as long as the available alternative messages that

	•	C	S_1	S_2	S_1C	S_2C	S_1S_2	S_1S_2C
Player A cheered	0	.43	0	0	.23	.23	0	.10
Player A hit every shot	0	0	0	0	0	0	.72	.28
Player A hit some shot	0	0	.33	.33	.09	.09	.10	.04
Player A hit some shot or cheered	0	.15	.28	.28	.08	.08	.09	.03
Player A hit some shot and cheered	0	0	0	0	.41	.41	0	.17
Player A hit every shot or cheered	0	.34	0	0	.19	.19	.20	.08
Player A hit every shot and cheered	0	0	0	0	0	0	0	1
0	1	0	0	0	0	0	0	0

Table 2: Inferences from nested scalar terms arising from competition among messages alone. (Introducing lexical uncertainty into the model only strengthens the basic patterns seen here.)

the speaker might have used is comprehensive in that it fully crosses the alternatives for *some* with the alternatives for *or*.

As table 2 shows, our model suffices to achieve this even with a fixed lexicon. For simplicity, we assume there are just two shots in the domain. Columns indicate the truth values of individual predicates: in S_1, Player A made the first shot, missed the second, and didn't cheer; in S_1S_2, Player A made every shot but didn't cheer; in C, Player A made no shots but cheered; in •, Player A made no shots and didn't cheer; and so forth. The crucial analytic step is to define the set of messages M so that it covers the space that Sauerland described. This suffices to capture the desired inferences in the probabilistic sense that our model provides: given *Player A hit some shot or cheered*, our pragmatic listener (13c) places most of its probability mass on worlds in which Player A only cheered (C) or made only some shots and did not cheer (S_1, S_2). We also see the expected scalar inferences from *some* and *or* when they appear independently: *Player A hit some shot and cheered* leads the listener away from states where both shots were made, and *Player A hit every shot or cheered* leads the listener away from the world verifying both conjuncts, S_1S_2C.

We obtained the results of table 2 using a uniform prior over states, but similar qualitative patterns would hold using a different prior specification. Likewise, allowing lexical refinements, as in the full version of our model, strengthens the relevant inferences without changing the qualitative pattern seen in table 2. For brevity we do not show this result, but readers are encouraged to try the simulations for themselves, using the code provided with this paper.

Let's now look at a larger and more complex scenario, one in which lexical uncertainty interacts with message competition to help reveal the potential of this model to capture embedded implicatures in ways that a fixed-lexicon version of the model cannot. In this scenario, there are two players. We resume our convention of referring to worlds using sequences like NN ('neither player scored'). The lexical items are *Player A*, *Player B*, *some*, *every*, *no*, *scored*, and *aced*. To start, we assume that, for all lexical items φ, $\mathcal{R}_c(\varphi) = \wp([\![\varphi]\!]) - \emptyset$. This creates an enormous space of lexica, and allows the full range of possible interactions between the refinements.

The listener inferences are summarized in table 3. For the most part, they seem aligned with the general view in the literature about how scalar terms interact in contexts like this. For instance, we predict that a proper name *P* will take on the exhaustified sense *only P*, as we would expect given the salience of *every*. In turn, *some* is interpreted as non-specific in virtue of the salience of

the two names, and it also leads to an SI due to the salience of *every*. Perhaps the most striking outcome is that the scalar inference from *scored* to not-aced remains in evidence not just with the proper names but also in the scope of the quantified subjects: the best-guess inference for *every player scored* is SS. These effects derive from interacting lexical uncertainty between the subjects and predicates.

Table 3 reveals some drawbacks to unfettered exploration of refinements, however. First, we might expect hearing *some player scored* to lead the listener to assume that the state was either NS or SN, corresponding to enrichment of both the subject ('not all players') and the predicate ('merely scored'). The current model does not achieve this. In addition, the row for *no player scored* is unintuitive. The best inference is NN, which is in line with the literal semantics, but it is striking that the states NS and SN have some positive probability. This arises because of interacting lexical uncertainty: there are lexica in the space in which *scored* is refined to exclude one of the players. In that case, the negative universal turns out to be true. Only a few lexica support this interaction, ensuring that it cannot become dominant, but it still seems worrisome.

This worry is a touchstone for revisiting an assumption of the model underlying table 3: that the lexical items can be refined in completely arbitrary ways. We take it to be one of the major lessons of neo-Gricean approaches that alternatives are contextually and lexically constrained. CFS's treatment of *ALT* reflects this lesson, as do our own sets of alternative messages M. Our handling of refinement allows us to incorporate such insights at the level of lexical uncertainty as well. This is not part of the neo-Gricean perspective as normally construed, but it's a natural step in the context of our model. Thus, it is worth seeing whether we can improve the picture in table 3 by encoding lexical scales in our grammar fragment.

We implement lexical scales in our model by constraining the refinement sets for several lexical items, as follows:[6]

(14) a. $\mathcal{R}_c(\textit{Player A}) = \{[\![\textit{Player A}]\!], [\![\textit{only Player A}]\!]\}$

b. $\mathcal{R}_c(\textit{Player B}) = \{[\![\textit{Player B}]\!], [\![\textit{only Player B}]\!]\}$

c. $\mathcal{R}_c(\textit{some}) = \{[\![\textit{some}]\!], [\![\textit{some and not all}]\!]\}$

d. $\mathcal{R}_c(\textit{no}) = \{[\![\textit{no}]\!]\}$

e. $\mathcal{R}_c(\textit{scored}) = \{[\![\textit{scored}]\!], [\![\textit{scored and didn't ace}]\!]\}$

f. $\mathcal{R}_c(\textit{aced}) = \{[\![\textit{aced}]\!]\}$

The results of working in this more constrained, neo-Gricean refinement space are given in table 4. The picture is mostly unchanged, except we now also achieve the target enrichment for *some player scored*, and the messiness surrounding *no player scored* is fully addressed. The one remaining potential concern about table 4 is that it predicts rather aggressive pragmatic enrichment of the scalar term in the scope of the negative quantifier. As we noted in section 2, it has long been assumed that weak scalar items in such environments fail to give rise to upper-bounding implicatures. Chemla & Spector (2011) address this question empirically, finding in their experiment low but non-negligible rates of local enrichment in negative environments. We too treat this as an empirical question; in section 6, we present evidence that local enrichments of this sort are indeed salient possibilities for humans.

[6] We define $[\![\textit{only Player A}]\!] = \{\langle w, Y\rangle : \{\mathtt{a}\} = \{x : \langle w, x\rangle \in Y\}\}$, and similarly for $[\![\textit{only Player B}]\!]$, not as a claim about natural language *only*, but rather just for the sake of the simulation.

	NN	NS	NA	SN	SS	SA	AN	AS	AA
Player A scored	0	0	0	.24	.19	.16	.18	.16	.07
Player A aced	0	0	0	0	0	0	.36	.30	.34
Player B scored	0	.24	.18	0	.19	.16	0	.16	.07
Player B aced	0	0	.36	0	0	.30	0	0	.34
some player scored	0	.14	.11	.14	.17	.14	.11	.14	.05
some player aced	0	0	.22	0	0	.19	.22	.19	.18
every player scored	0	0	0	0	.31	.27	0	.27	.14
every player aced	0	0	0	0	0	0	0	0	1
no player scored	.31	.14	.12	.14	.06	.05	.12	.05	.01
no player aced	.18	.19	.08	.19	.14	.06	.08	.06	0
0	.01	.01	.32	.01	.01	.15	.32	.15	0

Table 3: Enrichment in the largest space of refinements supported by this lexicon.

	NN	NS	NA	SN	SS	SA	AN	AS	AA
Player A scored	0	0	0	.45	.11	.22	.15	.05	.02
Player A aced	0	0	0	0	0	0	.42	.36	.22
Player B scored	0	.45	.15	0	.11	.05	0	.22	.02
Player B aced	0	0	.42	0	0	.36	0	0	.22
some player scored	0	.25	.09	.25	.06	.12	.09	.12	.01
some player aced	0	0	.24	0	0	.21	.24	.21	.11
every player scored	0	0	0	0	.61	.16	0	.16	.07
every player aced	0	0	0	0	0	0	0	0	1
no player scored	.61	0	.16	0	0	0	.16	0	.06
no player aced	.19	.17	.10	.17	.13	.07	.10	.07	0
0	.15	.13	.13	.13	.10	.09	.13	.09	.05

Table 4: Enrichment using the lexically-driven (neo-Gricean) refinement sets in (14).

5 Prior experimental work

The illustrative examples in the previous section begin to show that our compositional lexical uncertainty model naturally generates local enrichments. Thus, the question of whether listeners actually make such inferences is critical in judging the suitability of this model as a description of human reasoning. The present section reviews the prior literature in this area.

The pioneering paper is Geurts & Pouscoulous 2009. Their experiments 3 and 4 asked participants to provide truth-value judgments for sentences paired with abstract visual scenes consisting of shapes connected by lines. The target sentences included weak scalar terms in upward, downward, and non-monotone contexts, such as *exactly two of the squares are connected with some of the circles*, comparable in relevant respects to the examples reviewed in section 2 above. Geurts & Pouscoulous found only negligible rates of inferences consistent with local enrichment. These findings stimulated a number of responses commenting on the prevalence of local enrichment and its theoretical import (Ippolito 2010; Sauerland 2010). The two responses that are most

relevant for our purposes are those of Clifton & Dube (2010) and Chemla & Spector (2011).

Clifton & Dube (2010) argue that the experimental setting used by Geurts & Pouscoulous was prone to understating the rate of implicatures, and they sought to address this issue with a different experimental method. In their experiment, one trial consisted of presenting the participant with a sentence together with a set of visual scenes. The participant was instructed to choose the scene or scenes, if any, that he or she considered "best described by the sentence". They found that participants tended to chose the scene consistent with local enrichment. This method is a natural choice given a pragmatic model like ours, since it places participants in a role comparable to that of a listener agent. The particulars of the experimental method were criticized by Geurts & van Tiel (2013: §5.1) and van Tiel (2014), however, on the grounds that, for the examples involving monotone quantifiers, the inferences are better explained in terms of the typicality effects of the quantifiers involved (see also Degen & Tanenhaus 2015). Roughly speaking, the claim is that the typicality structure of *some A are B* favors situations in which just shy of half the A's are B's, and experimental designs (like Clifton & Dube's) that allow participants to express extra-truth-conditional preferences will be sensitive to this typicality structure. While we think that typicality is an important component of many implicatures and thus should ultimately be derived from a complete pragmatic model rather than considered a separate, overlaid factor,[7] we also see value in trying to neutralize its effects for purposes of studying local enrichment.

Chemla & Spector (2011) followed Geurts & Pouscoulous (2009) in asking participants to interpret quantified sentences in abstract geometric scenes, but they sought to simplify those scenes (see Geurts & van Tiel 2013: 31 for criticisms of this presumption), and they allowed subjects to provide graded truth-value judgments on a scale between 'Yes' and 'No'. The results were consistent with very high rates of local enrichment in upward and non-monotone environments, and even yielded suggestive evidence for local enrichment in downward monotone environments. These findings stand in stark contrast to those of Geurts & Pouscoulous (2009).

However, there are at least three features of Chemla and Spector's experimental design that might have exaggerated the rates of judgments consistent with local enrichment (Geurts & van Tiel 2013). First, the graded response categories mean that, for the monotone cases, typicality effects might have played a role. Second, the visual scenes were wheel-like displays in which lines extend from the vertex to the perimeter. There are potentially many ways this state can be drawn. Some might be more iconic than others, and some might create spurious patterns and salience contrasts that could affect linguistic inference in unmeasured ways. Third, Chemla & Spector used a within-subjects design: the individual participants judged every sentence in every context. Participants could thus have drawn comparisons across different conditions, creating opportunities for them to register comparative judgments involving the experimental contexts themselves, rather than relying solely on their linguistic intuitions.

We draw three major lessons from the above studies and debates. First, we should seek out simple, naturalistic stimuli. Previous experiments in this area have used abstract displays. Together with the inevitable complexity of the sentences involved, this choice seems likely to put cognitive demands on participants in ways that could affect the stability and reliability of the responses. Second, scalar response categories might encourage typicality inferences that could cloud

[7] Levinson's (2000) I-implicatures involve inferences from a general term or statement to one of its salient or prototypical subkinds. In the context of a generalized theory of scalar (partial-order) inference like that of Hirschberg (1985), this can be seen as a scalar inference guided by prior expectations.

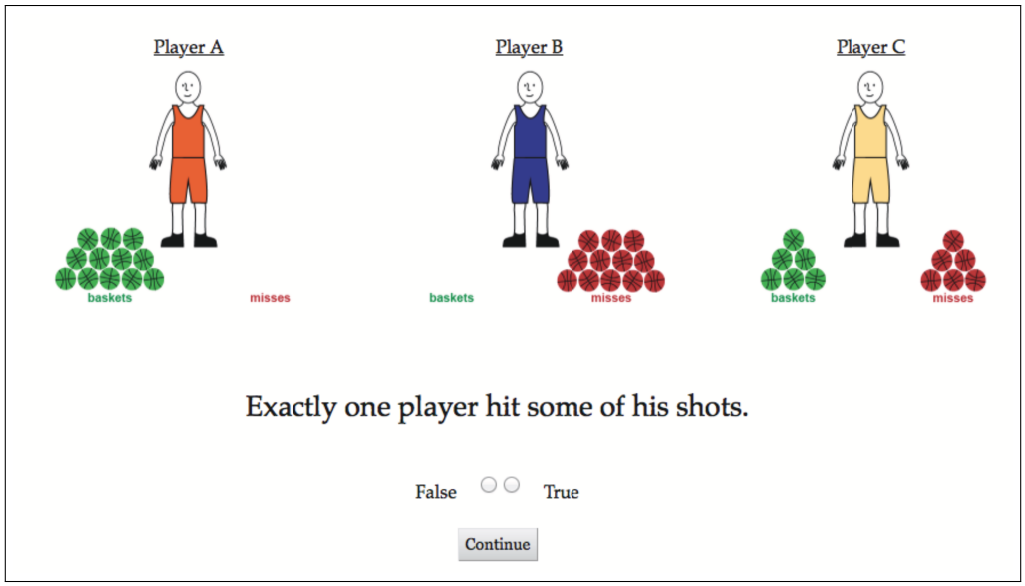

Figure 3: Experiment display.

the implicature picture; this might be a concern only for monotone environments, but we can hope to avoid the issue by restricting to just truth-value responses. Third, to the greatest extent possible, we should seek a design that supports analyses in which we can marginalize out the idiosyncrasies of particular displays, to avoid artifacts of salience or contrast that could stimulate responses that are consistent with implicature calculation without requiring such calculation.

6 Experiment: Scalars under quantifiers

We now present our main experiment involving *some* in quantified environments. We told participants that they were helping to train an automated sportscasting system and asked them to provide truth-value judgments about sentences in the context of displays like figure 3. This cover story was designed to ensure that implicatures are relevant, that is, worth calculating where available (Chemla & Spector 2011: §3.1; Clifton & Dube 2010). Our goal was to better understand the extent to which certain pragmatic inferences are available, so we sought out a scenario that would be maximally favorable to them. (For studies aimed at understanding the prevalence of implicatures, see Paris 1973; Hendriks et al. 2009; Degen 2015.)

6.1 Methods

6.1.1 Participants

The experiment had 800 participants, all recruited with Amazon's Mechanical Turk. No participants or responses were excluded.

6.1.2 Materials

We generated displays like those in figure 3. In each display, each of the three players, A, B, and C, has taken 12 basketball shots (a number small enough for visual display but outside of the subitizing range and thus less likely to introduce competitions from cardinal determiners like *three shots*; Degen & Tanenhaus 2015). The shots were divided into two piles, labeled 'baskets' (green) and 'misses' (red). For our target items, the player either made all 12 baskets (Player A in figure 3), missed all 12 baskets (Player B), or made 6 and missed 6 (Player C). The colors of the players' clothes were set randomly from a palette of 14 colors.

The target sentences describing the displays were defined as follows:

(15) $\left\{\begin{array}{l}\text{Every}\\ \text{Exactly one}\\ \text{No}\end{array}\right\}$ player hit $\left\{\begin{array}{l}\text{all}\\ \text{none}\\ \text{some}\end{array}\right\}$ of his shots.

Following previous studies, we put a bound pronoun in the embedded quantifier to try to ensure that the subject took scope over the object. The partitive forms seem likely to further encourage implicature calculation (Reed 1991; Grodner et al. 2010; Degen 2015). We chose the verb *hit* over the slightly less marked verb *make* to try to avoid the sense of 'make' as in 'take' (consistent with missing).

For the target items, there were ten different conditions, corresponding to the worlds in (16), in the notation we've been using to identify possible worlds.

(16) {NNN, NNS, NNA, NSS, NSA, NAA, SSS, SSA, SAA, AAA}

This is a subset of the full cross-product of the three outcomes N, S, and A in which player i always did at least as well as player $i+1$, going left to right. Our target sentences were all quantified, so we don't care about the outcome for any single player, meaning that we don't distinguish, e.g., NNS from NSN, allowing us to work with this smaller set of conditions. In the experiment, the 'order' of each world was randomized, so that, e.g., NSA appeared visually in each of its three orders approximately the same number of times. This randomization allows us to control for preferences in visual processing that might naturally make one position or linear ordering of player outcomes salient in unanticipated ways.

6.1.3 Procedure

After reading our consent form, participants were given the following cover story about "a basketball free-throw shooting competition between 3 players":

(17) We are trying to train an automated sportscasting system to generate color commentary on simple competitions. We'd like you to make judgments about the comments it generates. We'll use these ratings to train our system further.

After reading this cover story and some instructions, participants were presented with three training items, designed to ensure that participants understood the cover story, displays, and sentences. They then judged 32 sentences, divided into 9 target sentences and 23 fillers. The design was between-subjects: no experimental participant judged the same sentence twice. The order of presentation of the items was randomized.

Each sentence received a total of 800 responses. For the target sentences, each sentence–world pair received between 58 and 103 responses (mean 80); this variation resulted from randomization in the assignment of worlds to sentences.

Target sentences were presented below displays. Participants were asked to evaluate sentences as either true or false. In this sense, our participants acted as listeners who got to observe the speaker's state and assess whether the speaker accurately described that state with her utterance. We also conducted a variant of the experiment in which participants gave responses on a seven-point Likert scale ranging from 'Bad description' to 'Good description', to see whether this would reveal information about the quality of the report. These two versions of the experiment led to qualitatively identical outcomes. Appendix B reviews the details of the scalar-response version.

All the materials and response data for the experiment are available at the website for this paper.

6.2 Results

Figure 4 summarizes the responses by target sentence and the world in which it was evaluated. Overall, participants made judgments that accurately reflected whether sentences were true or false; accuracy was especially clear for the sentences in the first two columns, which do not admit pragmatic enrichment. For these cases, the responses were essentially categorical. This pattern suggests that our method is appropriate for measuring participants' interpretations.[8]

We now turn to the critical conditions, reporting significance levels for key theoretical comparisons based on the nonparametric Mann–Whitney U test. Responses for '*every…some*' (upper right) were consistent with the hypothesis that *some* is locally enriched in this condition. In particular, this sentence received the greatest percentage of 'True' responses in the SSS world. As we reviewed in section 2, in order to count as a complete report in this world, this sentence requires either local enrichment or a Gricean calculation with auxiliary premises. Worlds SSA and SAA received the next highest percentages of 'True' responses (lower than SSS, $p = 0.09$ and $p = 0.04$, respectively). Of all the literally true worlds for this condition, AAA received the lowest percentage of 'True' responses (lower than SSA and SAA; both at $p < 0.01$). Only a simple Gricean calculation is required to account for the higher rate of 'True' for SSA and SAA compared with AAA: in the latter world, the salient alternative *every player hit all of his shots* is a more complete description.

Nevertheless, '*every…some*' is not a strong test of the presence of local readings, since the entailment relations between the readings introduce some indeterminacy into the analysis. In particular, since the local enrichment entails the literal reading, we can't be sure whether the 'True' responses for SSS derive entirely from the availability of a local enrichment: a literal construal would suffice to make the sentence true. Furthermore, as discussed in section 2, '*every…some*' is of limited utility in distinguishing theoretical proposals anyway. It is the '*exactly one…some*' sentence that allows us to probe most confidently for local readings.

The response pattern for the critical item '*exactly one…some*' is given in the middle right of

[8]The only exception to this general pattern is the sentence *No player hit none of his shots* (bottom middle). The percentage of 'True' responses is lower than normal in all its true conditions and relatively high for NNN, where it is false on its literal construal. We hypothesize that this pattern reflects a negative concord construal, on which the embedded term is interpreted as equivalent to *any of his shots*, creating a meaning that is true only in NNN. Negative concord of this sort is productive in many dialects of English and understandable in all or nearly all of them. This likely created uncertainty about the intended meaning of the sentence, leading participants to disfavor it in general.

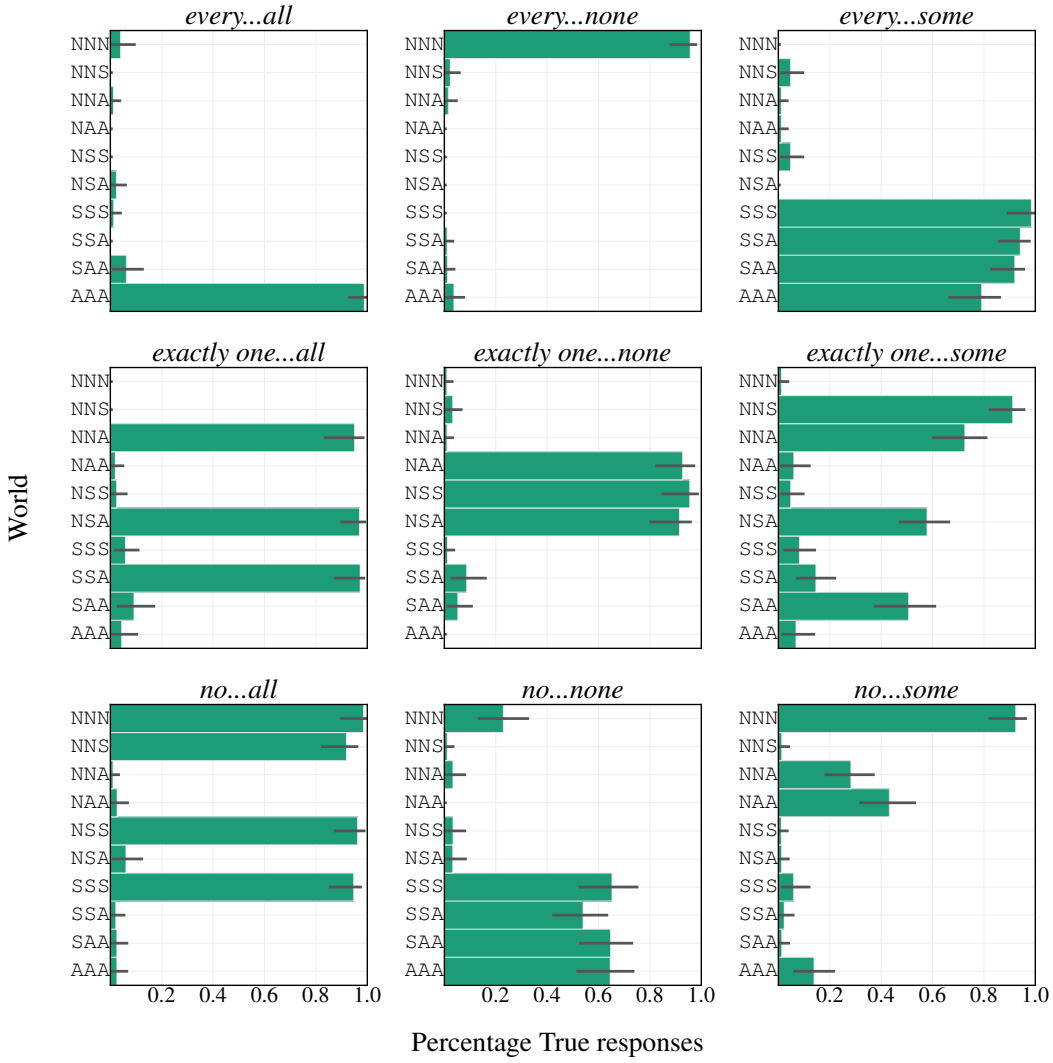

Figure 4: Mean truth-value judgment responses by sentence with bootstrapped 95% confidence intervals.

figure 4. The highest percentage of 'True' responses is for the NNS condition, where the sentence is true under its literal and local enrichment construals. However, it was also frequently judged true in the NSA and SAA worlds (both higher at $p < 0.001$ than in SSA, the world yielding the highest rating among those in which the sentence is false both literally and under all possible enrichments). For NSA and SAA, the sentence is true only with local enrichment (because two players hit at least some of their baskets in these worlds, ruling out the literal construal). We note also that its more strictly truth-conditional interpretation seems to be salient as well, as it was generally perceived to be true in the NNA condition.

Finally, the pattern for '*no…some*' also suggests a non-trivial amount of local enrichment: though NNN produced the highest rate of 'True' responses, indicating a preference for a literal construal, the 'True' rates for NNA, NAA, and AAA are consistently higher than for the most favored false worlds, NNS and NSA; all pairwise significance tests for the cross-product of {NNS, NSA} and {AAA, NNA, NAA} are significant at $p = 0.002$. These are the worlds in which no player hit only

some of his shots, the local enrichment. This finding seems consistent with the low but non-negligible rates of local enrichment that Chemla & Spector (2011: §4.4.4) report for this quantifier pair. One qualification we should add here is that our sentence is arguably somewhat unnatural in that it places *some*, a positive polarity item (Baker 1970; Israel 1996), in the scope of a negative quantifier. The binding relation between the subject and the pronoun *his* in the embedded phrase should force a surface-scope reading, but we can't rule out the possibility that participants might have found an inverse-scope construal ('some shots are such that no player hit them') that took the scalar term out of the scope of the negation. Alternatively, the marked nature of *some* in this position might have encouraged implicit prosodic focus, which would also likely increase the 'only some' construals.

We conclude from these responses that local enrichment is possible even in non-monotone environments, and that local enrichment might be available in downward-monotone environments as well. However, our concern is not only whether such readings are possible or impossible, but rather how accurately we can predict their availability on the basis of contextual and world knowledge. We turn now to the task of assessing the ability of the model presented in section 4 to match both the quantitative and qualitative patterns in our experimental data.

6.3 Model assessment

The pattern of data we observed is sufficiently precise and detailed that extracting its full theoretical value requires more than arbitrary statistical tests of simple null hypotheses — e.g., the null hypothesis that in the '*exactly one…some*' condition, ratings are the same for the worlds admitted by local enrichment as for those excluded under both global and locally-enriched interpretations. This and other such null hypotheses can be rejected with high confidence. Instead, to characterize the patterns of inference that give rise to the observed data, we use a model-comparison approach. In particular, we evaluate four related models that each embody different characterizations of linguistic meaning. By comparing these models, we can gain insights into the aspects of each that contribute to particular patterns of predictions.

Our assumption in this comparison is that our models provide a description of aggregate human behavior across individuals. In this sense, they are posed at Marr's (1982) 'computational theory' level. They instantiate claims about the task that our participants are attempting to perform and the assumptions that they use in performing it, but they are agnostic about the particular processes ('algorithms', in Marr's terminology) by which individuals perform it. In particular, the averaged binary responses that we take as our modeling target could come about via a number of routes. For example, individuals could be computing or approximating the full computations that we describe here and then stochastically making binary choices based on their estimates of the underlying probability distribution. Alternatively, they could also be pursuing any number of heuristic, approximate strategies that — when aggregated across individuals and trials — could yield a stable probability estimate. We remain agnostic about this issue here, but we note that a growing literature explores these different hypotheses linking computational-level models to psychological processes (e.g., Bonawitz et al. 2014; Griffiths et al. To appear; Sanborn et al. 2010; Vul et al. 2014).

For all the models, we take as given the literal semantics described in table 1, as well as the following features of the context:

(18) a. $D = \{\texttt{a}, \texttt{b}, \texttt{c}\}$

b. W = the set in (16)

c. $M = \left\{ Q(player)(hit(S(shot))) : \begin{array}{l} Q \in \{exactly\ one, every, no\}, \\ S \in \{every, no, some\} \end{array} \right\} \cup \{\mathbf{0}\}$

d. $C(\mathbf{0}) = 5$; $C(m) = 0$ for all $m \in M - \{\mathbf{0}\}$

e. Flat state prior: $P(w) = P(w')$ for all $w, w' \in W$

f. Flat lexicon prior: $P_{\mathbf{L}}(\mathcal{L}) = P_{\mathbf{L}}(\mathcal{L}')$ for all $\mathcal{L}, \mathcal{L}' \in \mathbf{L}$

The domain D and worlds W come directly from our human experiment. Similarly, the set of messages M corresponds to (15), with some adjustments to keep the logical grammar simple. We stipulate flat priors and even costs (other than the null message). As noted in section 4, we do not have empirical estimates for these values; though better fits to the human data can be achieved by adding assumptions about them, this risks overfitting to the particular data we have and thus overstating the true accuracy of the models. The value $C(\mathbf{0}) = 5$ was chosen arbitrarily; appendix A explores a wide range of values for it.

The models we consider are defined as follows:

(19) a. **Literal semantics**: the predicted values are the output of l_0, as in (13a), run on the messages defined in (18c).

b. **Fixed-lexicon pragmatics**: the predicted values are the output of the uncertainty listener (13c), but all the lexical items have only themselves as refinements, so that the reasoning is entirely in terms of the base lexicon in table 1.

c. **Unconstrained refinement**: the inferences of the uncertainty listener (13c) with $\mathcal{R}_c(some) = \wp([\![some]\!]) - \emptyset$

d. **Neo-Gricean refinement**: as in 'Unconstrained refinement', but with $\mathcal{R}_c(some) = \{[\![some]\!], [\![some\ and\ not\ all]\!]\}$, as in (14) of section 4.4, to extend neo-Gricean insights about alternatives into the lexical uncertainty aspects of our model.

These models represent a broad range of approaches to linguistic meaning. The first neglects pragmatics entirely (the model includes a contextual prior over states, but we define it as flat). The second is a version of the rational speech acts model of Frank & Goodman (2012) and Goodman & Stuhlmüller (2013), which has been shown to capture a broad range of SIs, but is known to be limited in its ability to derive manner implicatures and certain classes of embedded implicature (Bergen et al. 2012, 2014). The final two models are full versions of the one we presented in section 4. They represent opposite ends of the spectrum of non-trivial refinements. We saw in connection with table 3 and table 4 that there might be empirical value in greatly constraining the space of refinements.

We employ three methods of comparison: Pearson's correlation coefficient, which measures the linear correlation between the human responses and the model predictions; Spearman's rank correlation coefficient, which assesses how closely the human responses and model responses are aligned in terms of the rankings they predict; and the mean-squared error (MSE) of the model predictions as compared with the human responses, which summarizes the distance of the predictions from the human behavior. The use of these three measures allows us to assess which models

	Pearson		Spearman		MSE	
Literal semantics	.938	(.926 – .947)	.762	(.754 – .770)	.0065	(.0057 – .0075)
Fixed-lexicon pragmatics	.924	(.911 – .932)	.757	(.749 – .766)	.0079	(.0072 – .0090)
Unconstrained uncertainty	.945	(.936 – .950)	.794	(.767 – .820)	.0038	(.0035 – .0044)
Neo-Gricean uncertainty	.959	(.950 – .962)	.809	(.808 – .820)	.0034	(.0031 – .0040)

Table 5: Overall assessment with 95% confidence intervals obtained via non-parametric bootstrap over subjects.

best reproduce quantitative correspondence modulo arbitrary linear transformation (Pearson correlation), qualitative correspondence (Spearman correlation), and absolute fit between models and data. We find that the Spearman measure is often the most illuminating, since our fundamental goal is to reproduce the preference orderings revealed by the human responses. However, the three measures together yield a succinct multidimensional summary of how the models fare, and the same measures can be applied to particular target sentences to achieve more fine-grained insights.

Our model predictions are conditional probability distributions over states given messages, and hence constrained to be in the range $[0, 1]$ and to sum to 1. In contrast, our human responses are binary true/false judgments. To align these values, we rescale the human responses: if x^s is the 10-dimensional vector of percentage-true human responses for target sentence s, then each p^s is the vector of normalized values for that sentence, defined so that $p^s_i = x^s_i / \sum_{j=1}^{10} x^s_j$. This simply normalizes the responses into a conditional probability distribution over states given messages. The one noteworthy thing about this calculation is that, because it is done on a per-sentence basis, it is not a simple linear rescaling, and so it affects all of our assessment metrics when applied to multiple sentences at once. However, we regard it as the simplest viable linking hypothesis relating our model with our experimental data.

Figure 5 summarizes the models' predictions alongside the human responses. The predicted values are largely aligned for the examples without *some* in the object position. Where *some* occurs embedded, the models diverge in qualitative terms. For '*every*…*some*', the patterns are broadly similar, but only 'Neo-Gricean uncertainty' is able to mirror the preference ordering of responses seen in the human data. For '*exactly one*…*some*', only the two uncertainty models are able to predict local enrichment, in that only they assign high probability to the crucial worlds that are false on the literal construal: `NSA` and `SAA`. The 'Literal semantics' and 'Fixed-lexicon pragmatics' models are unable to predict the salience of these construals. Similarly, only the two uncertainty models predict '*none*…*some*' to have embedded enrichments leading to acceptability for `NNA`, `NAA`, and `AAA`. In broad strokes, we can say that 'Fixed-lexicon pragmatics' predicts only 'global' implicatures, those that CFS would obtain with unembedded exhaustification, whereas the two uncertainty models simulate embedded exhaustification (though without predicting it to be the most preferred option, in line with our human responses).

Table 5 summarizes our overall quantitative assessment. All of the correlations are extremely high, and the MSE values are extremely low. This is reassuring about the general utility of all of these models for predicting human judgments. In addition, the confidence intervals on the estimates are tight. We computed confidence in these estimates by a subject-wise non-parametric bootstrapping procedure, recomputing correlations for the same set of conditions, but with different simulated samples of participants. The resulting intervals reflect our confidence about estimates of

Figure 5: Analysis by target sentence, comparing model predictions with human responses.

these statistics for this particular set of experimental conditions.

Because of the high absolute values of all correlations, model comparison is important for interpretation. Two patterns stand out. First, 'Fixed-lexicon pragmatics' performs the least well overall. Since it has been shown to add substantial value in other areas of language and cognition, we conclude that its particular approach to enrichment is at odds with the patterns for embedded implicatures. The precise causes are hard to pinpoint, but the fact that our target implicatures are not always enrichments of the literal content is surely part of the problem. Second, neo-Gricean uncertainty achieves the best results across all three of our measures. Here again, this is consistent with our expectations based on the large illustrative example from section 4.4, where we saw that this constrained, lexically-driven approach to choosing refinements resulted in the best quantitative and qualitative pattern.

The overall analysis given in table 5 understates the value of both uncertainty models when it comes to the distribution of embedded implicatures. Our target sentences provide relatively little space for pragmatic enrichment; in figure 4, the left and middle columns essentially have only literal interpretations, leaving just the right column for our pragmatic models to shine. What's more, our qualitative review of figure 5 suggests that the right column examples reveal major distinctions. It's thus worth assessing them quantitatively in isolation. The results of such an assessment are in table 6. The most dramatic pattern is that the two fixed-lexicon models are unable to capture the patterns for embedded implicatures in the non-monotone and downward monotone environments. In contrast, both uncertainty models capture the patterns. These tight fits are evident in figure 5, and it is reassuring to see them reflected in our assessment measures.

It is striking that the literal model is competitive for '*every…some*'. This model does not distinguish among contexts in which the target sentence is true. Our participants only minimally distinguished among such readings, which makes sense in the context of a binary judgment task if we assume that the literal reading is accessible. However, the distinctions that do emerge from our experimental results align best with the preference-order predicted by the 'Neo-Gricean uncertainty' model, as revealed by the high Spearman coefficient.

Finally, it seems that neither uncertainty model is clearly superior to the other for these data: they are the top two models on all metrics, and are separated from each other by only a small amount. This suggests to us that we may have not yet found precisely the right approach to refinement. It is tempting to try additional refinement sets to find a single model that wins decisively for all the target examples. We are wary of doing this because, as noted above, it runs the risk of overfitting to our experimental responses; we could easily engineer our own success. However, this is nonetheless a fruitful avenue for future exploration if paired with additional experiments for further validation. Appendix A offers additional relevant findings.

Our model's performance is sensitive to the space of competitor messages, so it is worth asking how robust these findings are to changes in this area. We have found that the basic pattern is robust to a number of changes to the space of quantifiers. The only noteworthy finding we have to report in this regard is that allowing *only some* into object position has a major impact: while SSS remains the best-guess inference for the message '*every…some*' in this setting, '*exactly one…some*' and '*no…some*' effectively lose their embedded implicature readings. This makes intuitive sense given the nature of the model: if the speaker has the option to choose *only some of his shots*, and that form is equally costly, then surely her avoidance of that form in favor of *some of his shots* is a signal that she regards the local enrichment as infelicitous. As *only some* is made more costly, it becomes a less salient option, and embedded implicatures begin to reemerge.

	'every…some'			*'exactly one…some'*			*'no…some'*		
	P	S	MSE	P	S	MSE	P	S	MSE
Literal	.99	.86	.0002	.80	.70	.0180	.88	.52	.0346
Fixed-lexicon	.93	.85	.0027	.80	.70	.0179	.88	.52	.0346
Unconstrained	.88	.84	.0043	.98	.94	.0007	.76	.57	.0097
Neo-Gricean	.82	.88	.0087	.94	.87	.0036	.93	.89	.0028

Table 6: Assessment of crucial items. 'P' = 'Pearson'; 'S' = 'Spearman'.

7 Conclusion

With this paper, we sought a synthesis between Gricean accounts of pragmatic reasoning and grammar-driven ones like that of Chierchia et al. (2012). It seems to us inevitable that both grammar and interaction will play leading roles in the final theory of these phenomena; at some level, all participants in the debate acknowledge this. Our achievement is to unify the crucial components of these approaches in a single formal model that makes quantitative predictions.

The key components of the model we develop are compositional lexical uncertainty and recursive modeling of speaker and listener agents (Bergen et al. 2014). The lexical uncertainty property is in evidence in Chierchia et al.'s account as well, in the form of underspecified logical forms with context-dependent meanings. Our model has similar formal mechanisms but also offers an account of how discourse participants reason under this persistent linguistic uncertainty. This leads to an important conceptual point: not all underspecification has to be resolved in order for robust pragmatic enrichment to take place.

The recursive reasoning of our model is characteristic of both Gricean approaches and signaling systems approaches; our model shares formal properties of both but makes quantitative predictions of the sort that can be correlated with human preferences in communication. There are by now many models in the same family as ours (see, e.g., Camerer et al. 2004; Jäger 2012; Smith et al. 2013; Kao et al. 2014b; Jäger & Franke 2014), so further exploration is likely to yield an even more nuanced picture.

In addition, we saw that the space of refinements has a significant impact on the final predictions. It would thus be worthwhile to further explore different notions of refinement, seeking better fits with our own experimental patterns and then validating those conclusions in follow-up experiments using our experimental items, or applying the resulting models in new domains. For example, whereas refinement in the present model applies only to lexical entries, it could apply to phrases as well. Such phrasal refinements might be required to account for what Sauerland (2012, 2014) has called 'intermediate implicatures', where scalar strengthening seems to apply in between two (potentially non-monotonic) operators. However, study of the empirical distribution of such implicatures and the precise formal assumptions required to account for that distribution has only just begun. We have made publicly available all the data and code associated with this paper in an effort to encourage these and other new strands of theory development and quantitative assessment.

			$C(\mathbf{0})$	λ	k
Literal semantics	Pearson	.94			
	Spearman	.76			
	MSE	.0065			
Fixed lexicon pragmatics	Pearson	.93	1	.1	1
	Spearman	.76	0	.2	1
	MSE	.0069	1	.1	1
Unconstrained uncertainty	Pearson	.97	1	.1	1
	Spearman	.80	1	.1	1
	MSE	.0022	1	.1	1
Neo-Gricean uncertainty	Pearson	.98	1	.1	1
	Spearman	.81	1	.2	1
	MSE	.0018	1	.1	1

Table 7: Best models found in hyper-parameter exploration, as assessed against the binary-response experiment. The literal listener is not affected by any of the parameters explored.

A Parameter exploration

As we discussed in section 4.3, the definition of our model naturally suggests at least two extensions: (i) a temperature parameter λ modulating the speaker's inferences, and (ii) further iteration beyond the level of L. The full extended form of the model is defined as follows, again drawing on the objects and notational conventions established in section 4.3:

(20) a. $l_0(w \mid m, \mathcal{L}) \propto \mathcal{L}(m, w)P(w)$

b. $s_1(m \mid w, \mathcal{L}) \propto \exp\left(\lambda\left(\log l_0(w \mid m, \mathcal{L}) - C(m)\right)\right)$

c. $L_1(w \mid m) \propto P(w) \sum_{\mathcal{L}\in\mathbf{L}} P_{\mathbf{L}}(\mathcal{L}) s_1(m \mid w, \mathcal{L})$

d. $S_k(m \mid w) \propto \exp\left(\lambda\left(\log L_{k-1}(w \mid m) - C(m)\right)\right)$ (for $k > 1$)

e. $L_k(w \mid m) \propto S_k(m \mid w)P(w)$ (for $k > 1$)

From the perspective of this model, our decision to set $\lambda = 1$ and focus on L_1 might appear arbitrary. In addition, even from the perspective of our simpler model, our decision to fix the cost of the null message at 5 for all simulations and assessments was arbitrary. It is therefore worth exploring other settings for these hyper-parameters. To do this, we conducted a comprehensive grid search of the following values:

(21) a. λ: $[0.1, 2]$ in increments of .1, and $[3, 5]$ in increments of 1

b. L_k for $k \in \{1, 2, 3, 4, 5, 6\}$

c. $C(\mathbf{0}) \in \{0, 1, 2, 3, 4, 5, 6\}$

The grid search explores the full cross product of these values for each of our four models. For each setting, we conduct our standard model assessment against the data from our main (binary-response) experiment. Table 7 reports the best values for each of our four models, along with

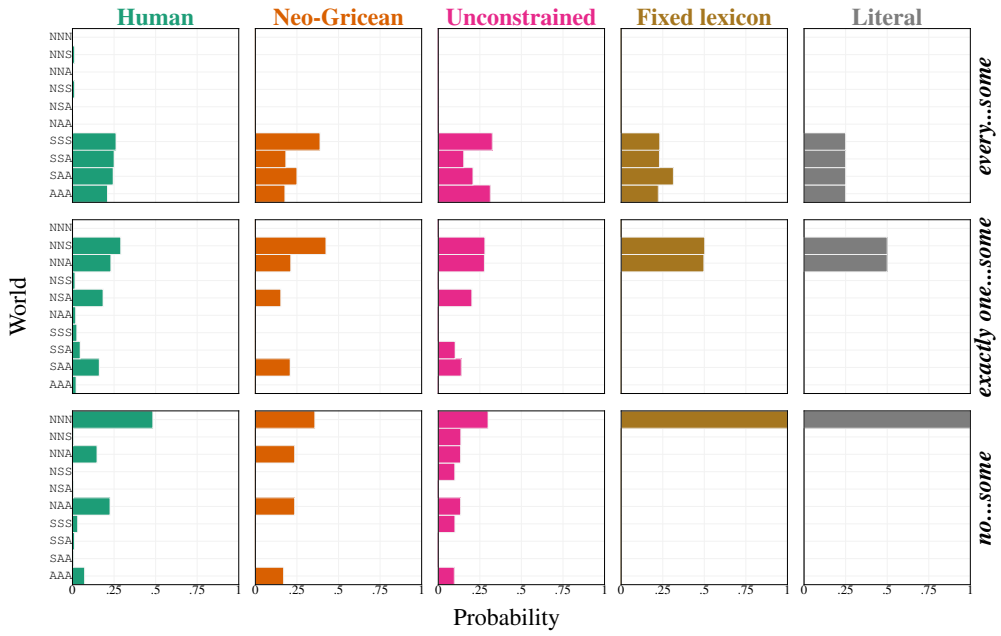

Figure 6: The crucial target sentences comparing the human data with L_1, using parameters in the range that seem to be nearly optimal for all of these models: $\lambda = 0.1$ and $C(\mathbf{0}) = 1$.

the minimal parameter settings that deliver those values. These results are consistent with our fundamental assessment of these models (section 6.3). Varying the cost of the null message has a relatively small impact on the outcomes, but the findings for the other two parameters may be relevant to broader discussions of *bounded* rationality in pragmatics. First, further iteration beyond L_1 is not necessary (Vogel et al. 2014). Second, the assumption in the main text that $\lambda = 1$, made primarily for clarity in deriving model predictions, does not provide the optimal fit to the experimental data: the value $\lambda = 0.1$ is slightly better. At lower values of λ, our listeners assume that speakers are paying little attention to the informativity of their messages, seeking only to be truthful (e.g., McMahan & Stone 2015). This is consistent with previous accounts according to which speakers are often unable to achieve ideal pragmatic calculations due to the the cognitive demands of production (Pechmann 1989; Levelt 1993; Engelhardt et al. 2006; Dale & Reiter 1995; van Deemter et al. 2012; Gatt et al. 2013). At the same time, the improvement is slight — compare table 7 to table 5 in the main text — and previous work has generally found that higher values of λ provide better predictions (for example, Kao et al. 2014a,b; Lassiter & Goodman 2015).

Figure 6 offers a finer-grained look at how these preferred settings affect outcomes for the crucial target items involving embedded *some*. The literal column is identical to the one in figure 5. The others are subtly different in ways that achieve a better match with the human data. For instance, the optimal parameters assign more probability to `AAA` in the '*no...some*' condition, which better matches the human responses. Overall, though, the contrasts between items are slightly dampened relative to the version of the model with $\lambda = 1$.

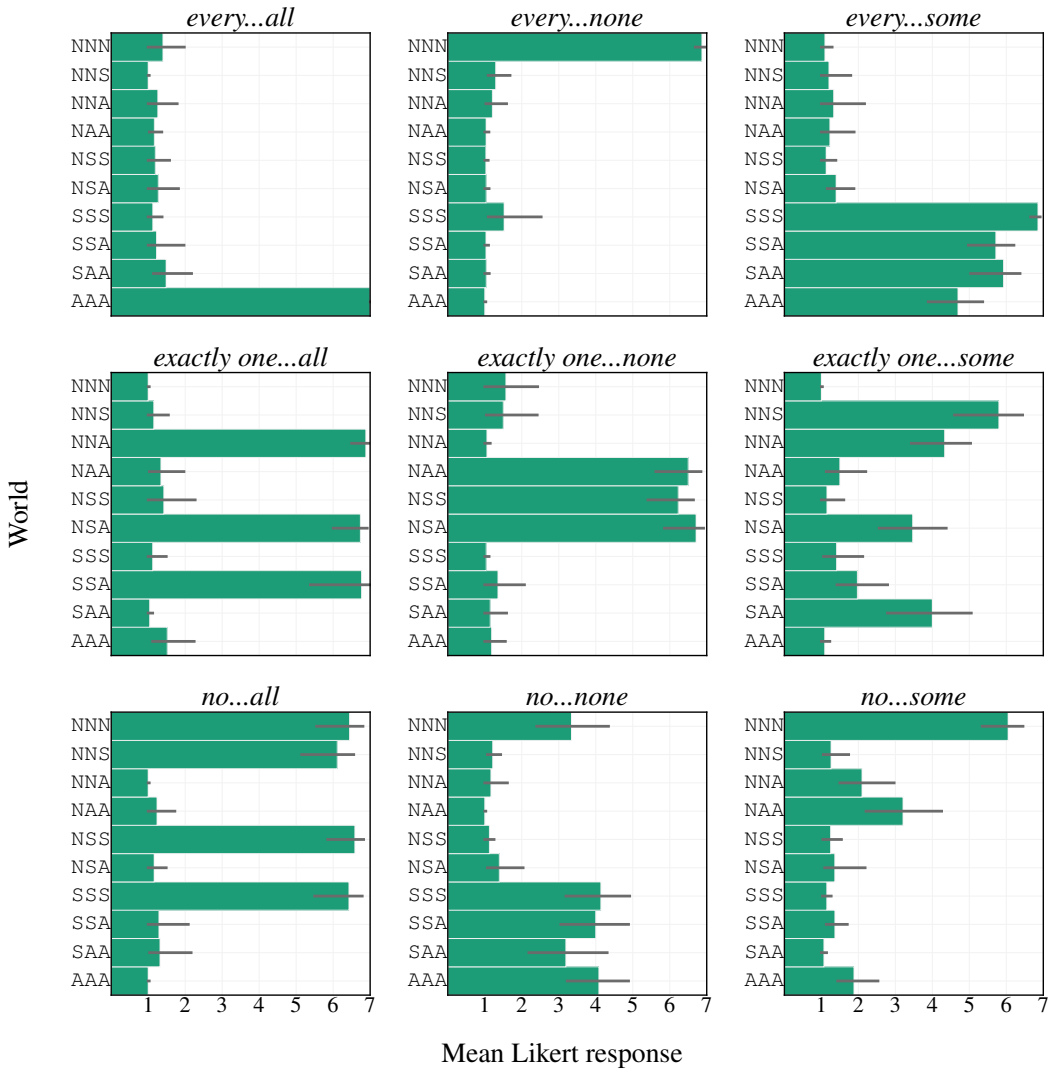

Figure 7: Likert-scale experimental results. Mean ratings by sentence with bootstrapped 95% confidence intervals.

B Likert-scale experiment

We conducted a version of the binary-response experiment discussed in section 6 using a Likert-scale for the response categories. Our rationale for using this scale was that it allows enough space for participants to both register a truth-value assessment and convey information about the quality of the report. This appendix reports briefly on this experiment. It yielded results identical in all important respects to those from our main experiment.

B.1 Methods

B.1.1 Participants

The experiment had 300 participants, all recruited with Amazon's Mechanical Turk. No participants or responses were excluded.

B.1.2 Materials

The displays were identical to those in figure 3, generated by the same procedures, but with the binary response categories replaced with a seven-point Likert scale ranging from 'Bad description' to 'Good description'. The target sentences were the ones in (15), and the conditions were as in (16). The same 23 fillers were used.

B.1.3 Procedure

After reading our consent form, participants were given the cover story in (17) with "judgments about the comments" replaced by "judgments about the quality of the comments". They completed the same three training items as were used in our main experiment. The design was again between-subjects. Each sentence received a total of 300 responses. For the target sentences, each sentence–world pair received between 19 and 44 responses (mean 30); this variation derives from our randomized procedure for assigning worlds to sentences.

B.2 Results

Figure 7 summarizes the responses by target sentence and world of evaluation. The results mirror those seen in figure 4 in all important respects. For our key theoretical comparisons, we again report significance levels using the nonparametric Mann–Whitney U test. In the '*every…some*' case, the highest ratings came in the SSS world. Worlds SSA and SAA received the next highest ratings (lower than SSS; both at $p < 0.001$). Of all the literally true worlds, AAA received the lowest rating (lower than SSA and SAA; both at $p < 0.05$). For the '*exactly one…some*' item, the highest ratings are again in the NNS condition, where it is true under its literal and locally enriched construals, but it also received high ratings in the two worlds where it is true only with local enrichment: NSA and SAA, which were both higher at $p < 0.05$ than in SSA, the world yielding the highest rating among those in which the sentence is false both literally and under all possible enrichments. As before, the strictly truth-conditional interpretation seems to be salient as well. Finally, we also find evidence for local enrichment under '*no…some*'. Condition NNN received the highest average ratings, suggesting a preference for a literal construal, but the ratings are high for the conditions requiring local enrichment: NNA, NAA, and AAA. The confidence intervals are wide, but a pooled comparison of {NNS, NSA} with {NNA, NAA, AAA} shows the latter set to be significantly higher-rated; $p = 0.006$.

B.3 Model assessment

Table 8 summarizes our model assessment. This assessment was done with identical settings and procedures to those reported in section 6.3, with one exception: since the minimal Likert value is 1,

	Pearson		Spearman		MSE	
Literal semantics	.935	(.910 – .947)	.756	(.742 – .764)	.0079	(.0065 – .0099)
Fixed-lexicon pragmatics	.920	(.894 – .932)	.751	(.736 – .759)	.0094	(.0080 – .0114)
Unconstrained uncertainty	.929	(.905 – .938)	.794	(.765 – .815)	.0052	(.0045 – .0067)
Neo-Gricean uncertainty	.950	(.927 – .956)	.805	(.795 – .812)	.0046	(.0038 – .0062)

Table 8: Overall assessment of the Likert-scale experiment with 95% confidence intervals obtained via by-subjects bootstrapping.

we subtract 1 from all scores when transforming them into the by-message normalized probability space of the model. Neo-Gricean uncertainty again emerges as the best model.

References

Alonso-Ovalle, Luis. 2008. Innocent exclusion in an alternative semantics. *Natural Language Semantics* 16(2). 115–128.

Bach, Kent. 1994. Conversational impliciture. *Mind and Language* 9(2). 124–162.

Bach, Kent. 2006. The top 10 misconceptions about implicature. In Betty Birner & Gregory Ward (eds.), *Drawing the boundaries of meaning: Neo-Gricean studies in pragmatics and semantics in honor of Laurence R. Horn*, 21–30. Amsterdam: John Benjamins.

Baker, C. L. 1970. Double negatives. *Linguistic Inquiry* 1(2). 169–186.

Beaver, David I. & Brady Zack Clark. 2008. *Sense and sensitivity: How focus determines meaning*. Oxford: Wiley-Blackwell.

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

Bergen, Leon, Roger Levy & Noah D. Goodman. 2014. Pragmatic reasoning through semantic inference. Ms., MIT, UCSD, and Stanford.

Blutner, Reinhard. 1998. Lexical pragmatics. *Journal of Semantics* 15(2). 115–162.

Bonawitz, Elizabeth, Stephanie Denison, Alison Gopnik & Thomas L Griffiths. 2014. Win-stay, lose-sample: A simple sequential algorithm for approximating Bayesian inference. *Cognitive Psychology* 74. 35–65.

Büring, Daniel & Katharina Hartmann. 2001. The syntax and semantics of focus-sensitive particles in German. *Natural Language and Linguistic Theory* 19(2). 229–281.

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

Chemla, Emmanuel. 2013. Apparent Hurford constraint obviations are based on scalar implicatures: An argument based on frequency counts. Ms. CNRS, ENS, LSCP Paris.

Chemla, Emmanuel & Benjamin Spector. 2011. Experimental evidence for embedded scalar implicatures. *Journal of Semantics* 28(3). 359–400.

Chierchia, Gennaro. 2004. Scalar implicatures, polarity phenomena, and the syntax/pragmatics

interface. In Adriana Belletti (ed.), *Structures and beyond: The cartography of syntactic structures*, vol. 3, 39–103. New York: Oxford University Press.

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

Chierchia, Gennaro, Danny Fox & Benjamin Spector. 2012. The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. In Maienborn et al. (2012) 2297–2332.

Clark, Eve V. & Herbert H. Clark. 1979. When nouns surface as verbs. *Language* 767–811.

Clark, Herbert H. 1997. Dogmas of understanding. *Discourse Processes* 23(3). 567–59.

Clifton, Charles Jr. & Chad Dube. 2010. Embedded implicatures observed: A comment on Geurts and Pouscoulous (2009). *Semantics and Pragmatics* 3(7). 1–13.

Dale, Robert & Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science* 19(2). 233–263.

van Deemter, Kees, Albert Gatt, Roger P.G. van Gompel & Emiel Krahmer. 2012. Toward a computational psycholinguistics of reference production. *Topics in Cognitive Science* 4(2). 166–183.

Degen, Judith. 2015. Investigating the distribution of *some* (but not *all*) implicatures using corpora and web-based methods. *Semantics and Pragmatics* 8(11). 1–55.

Degen, Judith & Michael K. Tanenhaus. 2015. Processing scalar implicature: A constraint-based approach. *Cognitive Science* 39(4). 667–710.

Engelhardt, Paul E., Karl G.D. Bailey & Fernanda Ferreira. 2006. Do speakers and listeners observe the Gricean maxim of quantity? *Journal of Memory and Language* 54(4). 554–573.

Fox, Danny. 2007. Free choice disjunction and the theory of scalar implicatures. In Sauerland & Stateva (2007) 71–120.

Fox, Danny. 2009. Too many alternatives: Density, symmetry, and other predicaments. In Tova Friedman & Edward Gibson (eds.), *Proceedings of Semantics and Linguistic Theory 17*, 89–111. Ithaca, NY: Cornell University.

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

Franke, Michael. 2009. *Signal to act: Game theory in pragmatics* ILLC Dissertation Series. Institute for Logic, Language and Computation, University of Amsterdam.

Gajewski, Jon. 2012. Innocent exclusion is not contradiction free. Ms., UConn.

Gatt, Albert, Roger P.G. van Gompel, Kees van Deemter & Emiel Krahmer. 2013. Are we Bayesian referring expression generators? In *Proceedings of the workshop on production of referring expressions: Bridging the gap between cognitive and computational approaches to reference*, Berlin.

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

Gazdar, Gerald. 1979b. A solution to the projection problem. In Choon-Kyu Oh & David A. Dinneen (eds.), *Syntax and semantics*, vol. 11: Presupposition, 57–89. New York: Academic Press.

Geurts, Bart. 2009. Scalar implicatures and local pragmatics. *Mind and Language* 24(1). 51–79.

Geurts, Bart. 2011. *Quantity implicatures*. Cambridge: Cambridge University Press.

Geurts, Bart & Nausicaa Pouscoulous. 2009. Embedded implicatures?!? *Semantics and Pragmatics* 2(4). 1–34.

Geurts, Bart & Bob van Tiel. 2013. Embedded scalars. *Semantics and Pragmatics* 6(9). 1–37.
Giles, Howard, Nikolas Coupland & Justine Coupland. 1991. Accommodation theory: Communication, context, and consequence. In Howard Giles, Nikolas Coupland & Justine Coupland (eds.), *Contexts of accommodation*, 1–68. Cambridge: Cambridge University Press.
Glucksberg, Sam. 2001. *Understanding figurative language: From metaphors to idioms*. Oxford University Press.
Goodman, Noah D. & Daniel Lassiter. 2015. Probabilistic semantics and pragmatics: Uncertainty in language and thought. In Shalom Lappin & Chris Fox (eds.), *The handbook of contemporary semantic theory*, Oxford: Wiley-Blackwell 2nd edn.
Goodman, Noah D. & Andreas Stuhlmüller. 2013. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science* 5(1). 173–184.
Grandy, Richard E. & Richard Warner. 2014. Paul grice. In Edward N. Zalta (ed.), *The stanford encyclopedia of philosophy*, Spring 2014 edn.
Grice, H. Paul. 1968. Utterer's meaning, sentence meaning, and word-meaning. *Foundations of Language* 4(3). 225–242.
Grice, H. Paul. 1975. Logic and conversation. In Peter Cole & Jerry Morgan (eds.), *Syntax and semantics*, vol. 3: Speech Acts, 43–58. New York: Academic Press.
Grice, H. Paul. 1989. *Studies in the way of words*. Cambridge, MA: Harvard University Press.
Griffiths, Thomas L, Falk Lieder, Noah D Goodman & Tom Griffiths. To appear. Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science* .
Grodner, Daniel J., Natalie M. Klein, Kathleen M. Carbary & Michael K. Tanenhaus. 2010. "Some," and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition* 116(1). 42–55.
Groenendijk, Jeroen & Martin Stokhof. 1984. *Studies in the semantics of questions and the pragmatics of answers*. Amsterdam: University of Amsterdam dissertation.
Hendriks, Petra, John Hoeks, Helen de Hoop, Irene Krammer, Erik-Jan Smits, Jennifer Spenader & Henriette de Swart. 2009. A large-scale investigation of scalar implicature. In Uli Sauerland & Kazuko Yatsushiro (eds.), *Semantics and pragmatics: From experiment to theory*, 30–50. Houndmills, Basingstoke, Hampshire: Palgrave Macmillan.
Hirschberg, Julia. 1985. *A theory of scalar implicature*. Philadelphia: University of Pennsylvania dissertation.
Horn, Laurence R. 1972. *On the semantic properties of logical operators in English*. Los Angeles: UCLA dissertation.
Horn, Laurence R. 1984. Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In Deborah Schiffrin (ed.), *Meaning, form, and use in context: Linguistic applications*, 11–42. Washington, D.C.: Georgetown University Press.
Hurford, James R. 1974. Exclusive or inclusive disjunction. *Foundations of Language* 11(3). 409–411.
Ippolito, Michela. 2010. Embedded implicatures? Remarks on the debate between globalist and localist theories. *Semantics and Pragmatics* 3(5). 1–15.
Israel, Michael. 1996. Polarity sensitivity as lexical semantics. *Linguistics and Philosophy* 19(6). 619–666.
de Jager, Tikitu & Robert van Rooij. 2007. Explaining quantity implicatures. In *Proceedings of the 11th conference on theoretical aspects of rationality and knowledge*, 193–202. New York:

ACM Digital Library.
Jäger, Gerhard. 2007. Game dynamics connects semantics and pragmatics. In Ahti-Veikko Pietarinen (ed.), *Game theory and linguistic meaning*, 89–102. Amsterdam: Elsevier.
Jäger, Gerhard. 2012. Game theory in semantics and pragmatics. In Maienborn et al. (2012) 2487–2425.
Jäger, Gerhard & Michael Franke. 2014. Pragmatic back-and-forth reasoning. In Salvatore Pistoia Reda (ed.), *Pragmatics, semantics and the case of scalar implicature*, 170–200. Houndmills, Basingstoke, Hampshire: Palgrave Macmillan.
Kao, Justine T., Leon Bergen & Noah D. Goodman. 2014a. Formalizing the pragmatics of metaphor understanding. In *Proceedings of the 36th annual meeting of the Cognitive Science Society*, 719–724. Wheat Ridge, CO: Cognitive Science Society.
Kao, Justine T., Jean Y. Wu, Leon Bergen & Noah D. Goodman. 2014b. Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences* 111(33). 12002–12007.
Lascarides, Alex & Ann Copestake. 1998. Pragmatics and word meaning. *Journal of Linguistics* 34(2). 387–414.
Lassiter, Daniel & Noah D. Goodman. 2013. Context, scale structure, and statistics in the interpretation of positive-form adjectives. In Todd Snider (ed.), *Proceedings of semantics and linguistic theory 23*, 587–610. Ithaca, NY: CLC Publications.
Lassiter, Daniel & Noah D. Goodman. 2015. Adjectival vagueness in a Bayesian model of interpretation. *Synthese* .
Levelt, Willem J.M. 1993. *Speaking: From intention to articulation*, vol. 1. MIT Press.
Levinson, Stephen C. 2000. *Presumptive meanings: The theory of generalized conversational implicature*. Cambridge, MA: MIT Press.
Lewis, David. 1969. *Convention*. Cambridge, MA: Harvard University Press.
Lewis, David. 1970. General semantics. *Synthese* 22(1). 18–67.
Magri, Giorgio. 2009. A theory of individual-level predicates based on blind mandatory scalar implicatures. *Natural Language Semantics* 17(3). 245–297.
Maienborn, Claudia, Klaus von Heusinger & Paul Portner (eds.). 2012. *Semantics: An international handbook of natural language meaning*, vol. 3. Berlin: Mouton de Gruyter.
Marr, David. 1982. *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: WH Freeman and Company.
McCawley, James D. 1978. Conversational implicature and the lexicon. In Peter Cole (ed.), *Syntax and semantics*, vol. 7: Pragmatics, 245–259. New York: Academic Press.
McMahan, Brian & Matthew Stone. 2015. A Bayesian model of grounded color semantics. *Transactions of the Association for Computational Linguistics* 3. 103–115.
Muskens, Reinhard. 1995. *Meaning and partiality*. Stanford, CA: CSLI/FoLLI.
Paris, Scott G. 1973. Comprehension of language connectives and propositional logical relationships. *Journal of Experimental Child Psychology* 16(2). 278–291.
Pechmann, Thomas. 1989. Incremental speech production and referential overspecification. *Linguistics* 27(1). 89–110.
Potts, Christopher & Roger Levy. 2015. Negotiating lexical uncertainty and speaker expertise with disjunction. In *Proceedings of the 41st annual meeting of the Berkeley Linguistics Society*, Berkeley, CA: BLS.
Reed, Ann M. 1991. On interpreting partitives. In Donna Jo Napoli & Judy Anne Kegl (eds.), *Bridges between psychology and linguistics: A Swarthmore festschrift for Lila Gleitman*, 207–

223. Hillsdale, NJ: Erlbaum.
Rooth, Mats. 1985. *Association with focus*. Amherst, MA: UMass Amherst dissertation.
Rooth, Mats. 1992. A theory of focus interpretation. *Natural Language Semantics* 1(1). 75–116.
Rooth, Mats. 1996. Focus. In Shalom Lappin (ed.), *Handbook of contemporary semantic theory*, 271–298. London: Blackwell.
Russell, Benjamin. 2006. Against grammatical computation of scalar implicatures. *Journal of Semantics* 23(4). 361–382.
Russell, Benjamin. 2012. *Probabilistic reasoning and the computation of scalar implicatures*. Providence, RI: Brown University dissertation.
Sanborn, Adam N., Thomas L. Griffiths & Daniel J. Navarro. 2010. Rational approximations to rational models: alternative algorithms for category learning. *Psychological Review* 117(4). 1144–1167.
Sauerland, Uli. 2001. On the computation of conversational implicatures. In Rachel Hastings, Brendan Jackson & Zsofia Zvolenszky (eds.), *Proceedings of Semantics and Linguistic Theory 11*, 388–403. Ithaca, NY: Cornell Linguistics Circle.
Sauerland, Uli. 2004. Scalar implicatures in complex sentences. *Linguistics and Philosophy* 27(3). 367–391.
Sauerland, Uli. 2010. Embedded implicatures and experimental constraints: A reply to Geurts & Pouscoulous and Chemla. *Semantics and Pragmatics* 3(2). 1–13.
Sauerland, Uli. 2012. The computation of scalar implicatures: Pragmatic, lexical or grammatical? *Language and Linguistics Compass.* 6(1). 36–49.
Sauerland, Uli. 2014. Intermediate scalar implicatures. In Salvatore Pistoia Reda (ed.), *Pragmatics, semantics and the case of scalar implicatures*, 72–98. Basingstoke: Palgrave MacMillan.
Sauerland, Uli & Penka Stateva (eds.). 2007. *Presupposition and implicature in compositional semantics*. Houndmills, Basingstoke, Hampshire: Palgrave Macmillan.
Schulz, Katrin & Robert van Rooij. 2006. Pragmatic meaning and non-monotonic reasoning: The case of exhaustive interpretation. *Linguistics and Philosophy* 29(2). 205–250.
Smith, Nathaniel J., Noah D. Goodman & Michael C. Frank. 2013. Learning and using language via recursive pragmatic reasoning about other agents. In *Advances in neural information processing systems 26*, 3039–3047.
Spector, Benjamin. 2007a. Aspects of the pragmatics of plural morphology. In Sauerland & Stateva (2007) 243–281.
Spector, Benjamin. 2007b. Scalar implicatures: Exhaustivity and Gricean reasoning. In Maria Aloni, Paul Dekker & Alastair Butler (eds.), *Questions in dynamic semantics*, 225–249. Amsterdam: Elsevier.
Sperber, Dan & Deirdre Wilson. 1995. *Relevance: Communication and cognition*. Oxford: Blackwell 2nd edn.
Stiller, Alex, Noah D. Goodman & Michael C. Frank. 2011. Ad-hoc scalar implicature in adults and children. In Laura Carlson, Christoph Hoelscher & Thomas F. Shipley (eds.), *Proceedings of the 33rd annual meeting of the Cognitive Science Society*, 2134–2139. Austin, TX: Cognitive Science Society.
Sutton, Richard S. & Andrew G. Barto. 1998. *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.
van Tiel, Bob. 2014. Embedded scalars and typicality. *Journal of Semantics* 31(2). 147–177.
Vogel, Adam, Andrés Gómez Emilsson, Michael C. Frank, Dan Jurafsky & Christopher Potts.

2014. Learning to reason pragmatically with cognitive limitations. In *Proceedings of the 36th annual meeting of the Cognitive Science Society*, 3055–3060. Wheat Ridge, CO: Cognitive Science Society.

Vul, Edward, Noah Goodman, Thomas L Griffiths & Joshua B Tenenbaum. 2014. One and done? Optimal decisions from very few samples. *Cognitive Science* 38(4). 599–637.

Wilson, Dierdre & Robyn Carston. 2007. A unitary approach to lexical pragmatics: Relevance, inference and ad hoc concepts. In Noel Burton-Roberts (ed.), *Pragmatics*, 230–259. Basingstoke and New York: Palgrave Macmillan.

Embedded implicatures as pragmatic inferences under compositional lexical uncertainty*

Christopher Potts, Daniel Lassiter, Roger Levy, and Michael C. Frank

August 29, 2015

Abstract

How do comprehenders reason about pragmatically ambiguous scalar terms like *some* in complex syntactic contexts? In many pragmatic theories of conversational implicature, local exhaustification of such terms ('only some') is predicted to be difficult or impossible if the result does not entail the literal meaning, whereas grammatical accounts predict such construals to be robustly available. Recent experimental evidence supports the salience of these local enrichments, but the grammatical theories that have been argued to account for this evidence do not provide explicit mechanisms for weighting such construals against others. We propose a probabilistic model that combines previous work on pragmatic inference under 'lexical uncertainty' with a more detailed model of compositional semantics. We show that this model makes accurate predictions about new experimental data on embedded implicatures in both non-monotonic and downward-entailing semantic contexts. In addition, the model's predictions can be improved by the incorporation of neo-Gricean hypotheses about lexical alternatives. This work thus contributes to a synthesis of grammatical and probabilistic views on pragmatic inference.

1 Conversational implicature: Interacting with grammar

The linguistic forms that discourse participants exchange with each other routinely underrepresent the speaker's intended message and underdetermine the listener's inferences. Grice (1975) famously provided a philosophical framework for understanding the driving forces behind such pragmatic enrichment. At the heart of this framework are **conversational implicatures**: social, cognitively complex meanings that discourse participants create jointly in interaction.

Perhaps the best-studied examples of language users going beyond the literal semantics involve weak terms like *some* being strengthened to exclude their communicatively stronger alternatives, giving rise to construals like 'some and not all' or 'only some'. Such inferences are often called **scalar conversational implicatures** (SIs), and they are widely assumed to arise via the same social inferencing mechanisms that are at work in other implicatures. However, this assumption has always been controversial. Even Grice suggested that SIs might be closer to the grammar than other implicatures (p. 56; see also Levinson 2000; Sperber & Wilson 1995; Bach 2006), and

*All the data and code used in this paper are available at `https://github.com/cgpotts/pypragmods`

recent grammar-driven accounts are framed in direct opposition to an implicature analysis. For example, Chierchia et al. (2012: 2316) write, "the facts suggest that SIs are not pragmatic in nature but arise, instead, as a consequence of semantic or syntactic mechanisms". The ensuing debates have stimulated new insights, pushing researchers to identify and evaluate previously unnoticed consequences of the two broad positions.

Much of the debate between Gricean and grammar-driven accounts has centered around what we informally called **embedded implicatures** — cases where a pragmatically enriched interpretation seems to be incorporated into the compositional semantics. Such readings seem initially to demand implicature-enriched semantic representations. However, many of the relevant examples have received straightforward Gricean accounts in which semantic content and contextual assumptions interact to yield global implicatures that are meaning-equivalent to interpretations that would derive from local pragmatic enrichment (Russell 2006; Geurts 2009, 2011). This reduces the power of such examples to decide in favor of one side or the other.

Geurts & Pouscoulous (2009) and Chemla & Spector (2011) study weak scalar terms in a wide range of quantificational environments. They show that many of the attested listener inferences concerning such terms are amenable to Gricean treatments based on implicature calculation, with no need for such calculations to intrude on the semantics (see especially Geurts & Pouscoulous 2009: §8 and Chemla & Spector 2011: 361). However, they identify a class of examples that, if attested, would not admit of such a treatment: scalar terms in the scope of non-monotone quantifiers, as in *exactly one player hit some of his shots*. In such cases, exhaustification of the embedded quantifier (. . . *some but not all of his shots*) does not entail the literal meaning, whereas the Gricean implicature analysis of scalar terms can only strengthen literal meanings. Geurts & Pouscoulous's experiments fail to support enrichment in such contexts, whereas Chemla & Spector's suggest that it is possible. A number of recent papers have sought to make sense of these conflicting results (Clifton & Dube 2010; Geurts & van Tiel 2013; van Tiel 2014).

In this paper, we reproduce the central qualitative result of Chemla & Spector (2011) using more naturalistic experimental stimuli, a fully randomized between-subjects design to avoid unwanted inferences across critical items (Geurts & van Tiel 2013), and a more direct method of interpreting participants' responses. Like Chemla & Spector, we find that scalar terms in non-monotone environments support implicature inferences (though these seem not to be the preferred or most salient construals). In our view, this evidence points to an important role for compositional semantics in understanding implicatures.

To describe the complementary roles of grammar and pragmatics in embedded implicatures, we propose a model that both embraces the compositional insights of Chierchia et al. and characterizes how people arrive at such construals. This model is in the tradition of **rational speech act** models (Frank & Goodman 2012; Goodman & Stuhlmüller 2013) and **iterated best response** models (Franke 2009; Jäger 2012), and is a direct extension of the **compositional lexical uncertainty** model of Bergen et al. (2012) and Bergen et al. (2014). The model accounts for how discourse participants coordinate on the right logical forms (implicature-rich or not), seeking to retain the insights of Gricean accounts while paying close attention to the details of semantic composition.

We show that our model not only captures the qualitative pattern of implicature behaviors that Chemla & Spector found, but also makes quantitative predictions that are highly correlated with people's actual inferential behavior in context. In addition, we present evidence that these correlations can be improved if the set of refinements is lexically constrained, in keeping with broadly

neo-Gricean views of SIs (Horn 1972; Gazdar 1979a,b; Schulz & van Rooij 2006), though the precise nature of the true refinements remains a challenging open question. Our results suggest that the full theory of implicature depends substantively on the fine details of semantic composition *and* broader considerations of rational interaction. This is perhaps a departure from Grice's (1975) particular conception of pragmatic meaning, but it is well-aligned with his general theory of meaning and intention (Grice 1968, 1989; Grandy & Warner 2014). In view of our experimental results, the chief advantage of our model is that it makes quantitative predictions that are easily and rigorously linked with our human response patterns. In other words, the model makes predictions not only about which pragmatic inferences are possible but also about how likely those inferences are.

Our broader position is that grammar-driven accounts and Gricean accounts are not in opposition, but rather offer complementary insights. When communicating in natural languages, people are relying on linguistic conventions to try to identify and convey each other's intentions. All sides in the debate acknowledge this mix of grammatical and interactional factors. Grice's (1975) definition of conversational implicature is interactional, but his maxim of manner embraces a role for linguistic form. By introducing additional devices such as Horn scales, Neo-Griceans expand this role into areas Grice addressed with the maxims of quantity, quality, and relevance. Sperber & Wilson (1995) and Bach (1994) characterize many kinds of pragmatic enrichment as inferences about logical forms. And Chierchia et al. (2012) invoke pragmatic pressures to explain how speakers and listeners coordinate on whether to posit implicature-rich logical forms or more literal ones. Thus, there is substantially more consensus than the rhetoric often suggests.

2 Implicature, enrichment, and embedding

In this section, we describe embedded implicatures, seeking to identify the special theoretical challenges they pose. Under Grice's (1975) original definition, conversational implicature is an act of social cognition. The original definition is somewhat underspecified, and fleshing it out into a precise formulation is challenging (Hirschberg 1985), but the guiding idea seems clear. The listener assumes that the speaker is cooperative in the Gricean sense of rational interaction. However, the listener is confronted with an utterance U with content p that meets this assumption only if certain additional conditions are met. The listener can resolve this tension by positing that these conditions are in fact met; in many (but not all) cases, this means inferring that the speaker intended for the listener to infer the truth of a different but related proposition q. By this reasoning, the listener is able to reconcile the observation that the speaker chose to utter U with the assumption that the speaker is communicating cooperatively.

In the current work, we do not try to make the above description more rigorous. The model that we develop does not depend on an independently formulated definition of implicature, but rather seeks to derive such meanings from more basic considerations about how speakers and listeners reason about each other whenever they interact. Similarly, the model of Chierchia et al. (2012) is noncommittal about the reality of conversational implicatures per se. In that model, 'conversational implicature' can be seen as an informal label for a certain class of logical forms, rather than a conceptual primitive (see section 3 of this paper). With this in mind, we use the notion of conversational implicature only to articulate the central empirical focus of this paper — embedded scalar terms — and the associated challenges for formal pragmatic accounts.

On the classic Gricean account, SIs arise when the imperative 'Be as informative as is required'

(a subclause of the maxim of quantity) is in tension with another pragmatic pressure related to cooperative communication. The opposing force can take many forms, for example, relating to considerations of politeness, discretion, or secrecy, but it is usually attributed to the maxim of quality, which instructs speakers to say only what they have strong positive evidence for. For instance, imagine a sportscaster who has observed the outcome of a single round of a basketball tournament and is reporting on it as news. If the sportscaster says (1), then she will likely implicate that Player A did not make all of his shots.

(1) Player A hit some of his shots.

The SI follows from a straightforward application of the above ideas. We assume that the sportscaster is cooperative in the Gricean sense, and knowledgeable and forthcoming about the events. Why, then, did she opt for a weak statement like *Player A hit some of his shots* when a stronger statement like *Player A hit all of his shots* is available and would have been more informative? If knowledge is the only relevant consideration, it must be that she was prevented from using this stronger form because she does not know it to be true. Together with our assumption that she observed the full outcome, she can lack knowledge of this proposition only because it is false, leading to the implicated meaning that Player A did not hit all of his shots. In this way, a listener can enrich the speaker's message.

To make this example more concrete, suppose that we have two players, A and B, and that we care (for present purposes) only about whether each of them hit none, some but not all, or all of his shots. We can identify these (equivalence classes of) possible worlds with labels like `NA`, which means that Player A hit none of his shots and Player B hit all of his shots, and `SS`, which means that both players hit some but not all of their shots. There are $3^2 = 9$ such worlds. The literal semantics of (1) in this context is the proposition given in (2b). Our hypothesized implicature is (2c), the proposition that Player A did not hit all of his shots. The intersection of these two meanings delivers the communicated meaning, (2d).

(2)

a. Worlds:	`NN NS NA SN SS SA AN AS AA`	
b. Literal:	`SN SS SA AN AS AA`	'at least some'
c. Implicature:	`NN NS NA SN SS SA`	'not all'
d. Communicated:	`SN SS SA`	'only some'

There are many proposals for how to formalize this reasoning. The common theme running through all of them is that the implicature is accessible because it is an enrichment that strictly entails the original literal content — in this example, because the utterance's literal meaning and the implicature are combined into a stronger meaning by intersection. In Grice's terms, a general claim is further restricted by the interaction of quantity and quality.

The above reasoning extends to examples like (3), in which *some* is in the scope of a universal quantifier, though additional assumptions must be brought in to achieve a comparable implicature.

(3) Every player hit some of his shots.

Consider the potential enrichment of this sentence to convey that every player hit some but not all of his shots. This seems comparable to the construal we derived for (1), but it requires more assumptions. If we take the implicature to be the negation of the stronger alternative *every player hit all of his shots*, then the reasoning proceeds as in the first four lines of (4), which takes us to a

meaning (4d) that is consistent with one or the other of the players (but not both) having hit all of his shots. To arrive at the target meaning (every player hit some but not all of his shots), we must further assume an auxiliary premise beyond that required for (1). One example of such a premise is that of uniform outcomes (4e); there are many others that will do the job (Spector 2007b).

(4)	a.	Worlds:	NN	NS	NA	SN	SS	SA	AN	AS	AA	
	b.	Literal:					SS	SA		AS	AA	'all hit at least some'
	c.	Implicature:	NN	NS	NA	SN	SS	SA	AN	AS		'not all hit all'
	d.	Result:					SS	SA		AS		'all hit some; not all hit all'
	e.	Aux. premise:	NN				SS				AA	'uniform outcomes'
	f.	Communicated:					SS					'all hit only some'

Though the need for an auxiliary premise is a noteworthy complication, it seems within the bounds of a Gricean account, and auxiliary premises like these might be independently justified (Russell 2006). As in the previous example, the communicated meaning is an enrichment of the literal content, and Gricean pressures and contextual assumptions deliver the stronger meaning. Geurts & Pouscoulous (2009) and Chemla & Spector (2011) home in on this common theme in scalar implicature calculation and use it to probe the scope and adequacy of the Gricean implicature framework. Examples like (5) are central to their discussions. This is a minimal variant of (3) with the subject universal determiner *every* replaced by *exactly one*.

(5) Exactly one player hit some of his shots.

Many people have the intuition that (5) can be used to describe a situation in which there is exactly one player who scored some but not all of his shots, which is consistent with some players having scored all of their shots. The reading is easy to characterize intuitively: one imagines that *some of his shots* has been locally enriched to *some but not all of his shots*, and that this enriched meaning is the semantic argument to the subject quantifier. What makes this reading notably different from, e.g., (3) is that it does not entail the literal reading, as we see in (6). The literal semantics is the proposition in (6b), whereas the content of the *. . . some but not all of his shots* ('Local') construal is (6c), which merely overlaps with it.

(6)	a.	Worlds:	NN	NS	NA	SN	SS	SA	AN	AS	AA	
	b.	Literal:		NS	NA	SN			AN			'exactly one hit at least some'
	c.	Local:		NS		SN		SA		AS		'exactly one hit only some'

Any theory in which enriched scalar interpretations are always generated by intersection, as they are in classical Gricean and neo-Gricean accounts, will fail to arrive at (6c). Such theories head inexorably toward a refinement that excludes NA and AN, but they are essentially incapable of 'introducing' SA and AS. If such construals are possible, they must arise from other mechanisms.

The issue is even clearer when a scalar term is in the scope of a downward-monotone operator like *no*, as in *no player hit some of his shots*. In such cases, the embedded enrichment creates a meaning that is strictly entailed by (i.e., weaker than) the literal meaning:

(7)	a.	Worlds:	NN	NS	NA	SN	SS	SA	AN	AS	AA	
	b.	Literal:	NN									'none hit some'
	c.	Local:	NN		NA				AN		AA	'none hit only some'

Gricean theories predict that the 'local' enrichment of *some* to *only some* is unavailable as an implicature inference here, either because of the way pragmatic pressures interact or because *some* is held to be the strongest member of its scale in negative environments, leaving no room for further enrichment. Grammar-driven approaches have tended to agree with the basic empirical assumption, arguing that local enrichment is blocked in environments where it would strictly weaken the literal content (Chierchia 2006).

The empirical evidence is mixed but seems to support the accessibility of these local interpretations. Modifying an earlier design by Geurts & Pouscoulous (2009), Chemla & Spector used displays involving geometric patterns to assess whether interpreters could access local-enrichment readings of scalar terms in the scope of non-monotone and downward-monotone operators. Their findings suggest that local enrichment readings are available in both contexts, especially non-monotone ones. Skeptics of local enrichment have found grounds for challenging Chemla & Spector's findings (see section 5), but we believe that the theoretical challenges posed by embedded implicatures are real. In section 6, we describe a new experiment that reproduces the core qualitative findings of Chemla & Spector's studies.

3 CFS's grammar-driven model

This section briefly reviews the grammar-driven model of Chierchia et al. (2012) (henceforth CFS). The approach is inspired by those of Chierchia (2004), Spector (2007a), and Fox (2007, 2009). There are two central pieces to the account: a generally available function *ALT* that maps words and phrases to their alternatives, and a covert exhaustification operator *O*.

For *ALT*, the relevant notion of alternative is familiar from theories of questions and focus (Groenendijk & Stokhof 1984; Rooth 1985, 1992): we can assume, as a default, that the alternatives for an expression φ is some subset of the items in the same type-theoretic denotation domain as $[\![\varphi]\!]$, the meaning of φ. The precise value of the function *ALT* is context-dependent, and discourse participants are presumed to coordinate on it, just as they coordinate on the meanings of deictic or discourse-bound pronouns, elided phrases, and other pragmatically controlled free variables.

The effect of applying the basic exhaustification operator *O* to an expression φ in the context of a given *ALT* is shown in (8) (Spector 2007a; Fox 2007, 2009; Magri 2009; Chierchia et al. 2012).[1]

(8) $O_{ALT}(\varphi) = [\![\varphi]\!] \sqcap \bigsqcap \{-q : q \in ALT(\varphi) \wedge [\![\varphi]\!] \not\sqsubseteq q\}$

The *O* operator maps an expression φ to one that entails $[\![\varphi]\!]$ and excludes the denotations of expressions in $ALT(\varphi)$ that are not strictly weaker than $[\![\varphi]\!]$. When dealing with truth-functional expressions, we can regard $\sqcap$ as boolean conjunction and $\sqsubseteq$ as a material conditional, but the definition should be thought of as broad enough to include any kind of partial ordering (Hirschberg 1985: §4).

Part of the case for a grammar-driven view is that it uses pieces of semantic theory that are independently needed. In particular, exhaustification is at the heart of Groenendijk & Stokhof's (1984) theory of questions and their answers. The above operator is a common proposal for the meaning of *only* (for discussion: Rooth 1996; Büring & Hartmann 2001; Beaver & Clark 2008). Schulz & van Rooij

[1] This is not the operator that CFS ultimately favor, since it requires some implicit restrictions on allowable *ALT* functions in order to get the right inferences. The final version has the same form as (8) but further restricts *ALT*.

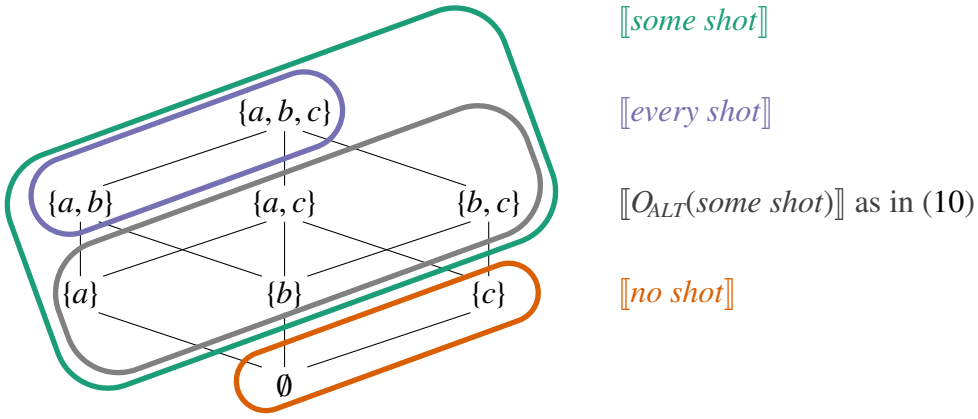

Figure 1: Given a domain $\{a, b, c\}$ with $[\![shot]\!] = \{a, b\}$, $[\![some\ shot]\!]$ is equal to the set of sets in the green box, $[\![every\ shot]\!]$ to the set of sets in the purple box, and $[\![no\ shot]\!]$ to the set of sets in the orange box. If *ALT*(*some shot*) contains $[\![every\ shot]\!]$, then *some shot* is refined to exclude the purple subset.

(2006) use exhaustification for implicature calculation (see also de Jager & van Rooij 2007). (For critical discussion, see Alonso-Ovalle 2008 and Gajewski 2012.) While CFS are cautious about making direction connections between *O* and these other phenomena (p. 2304), the correspondences are nonetheless noteworthy.

Those are the technical pieces. The proposal can then be summarized easily: *O* operators can optionally appear anywhere in the logical form of a sentence, perhaps subject to additional restrictions relating both to the comparative strength of the resulting logical form and to general pragmatic assumptions about the current conversational goals (see CFS: §4.6). To see the effects that this could have, let's return to the examples involving *some* that we reviewed in section 2. Simplifying slightly, let's suppose that *some shot* denotes the set of sets in (9) — the set of all sets *Y* that have a non-empty intersection with the set of shots.

(9) $[\![some\ shot]\!] = \{Y : [\![shot]\!] \cap Y \neq \emptyset\}$

Consider a domain of three entities $\{a, b, c\}$, and assume that $[\![shot]\!] = \{a, b\}$. Then the above is equivalent to the set of sets contained in the green box in figure 1. Now suppose that *ALT*(*some shot*) is defined as follows:

(10) $ALT(some\ shot) = \{[\![some\ shot]\!], [\![every\ shot]\!], [\![no\ shot]\!]\}$

a. $[\![some\ shot]\!]$ as in (9) (green circle in figure 1)

b. $[\![every\ shot]\!] = \{Y : [\![shot]\!] \subseteq Y\}$ (purple circle in figure 1)

c. $[\![no\ shot]\!] = \{Y : [\![shot]\!] \cap Y = \emptyset\}$ (orange circle in figure 1)

The presence of $[\![some\ shot]\!]$ has no effect because it is identical to the input. Similarly, all quantifiers that are weaker than the input have no effect if included in the *ALT* set. The presence of $[\![no\ shot]\!]$ has no effect because it contradicts the input, so its complement is weaker than the input. The presence of $[\![every\ shot]\!]$ will, though, be meaningful, as long as we assume that $[\![shot]\!] \neq \emptyset$. In that case, $O_{ALT}(some\ shot)$ will denote the subset in gray in figure 1. This is equivalent to the

intersection of $[\![\textit{some shot}]\!]$ and the complement of $[\![\textit{every shot}]\!]$ in the power set of the domain. In other words, it expresses *some and not all*, the intuitively implicature-rich interpretation.

Because O_{ALT} is embeddable, syntactic constituents like O_{ALT}(*some shot*) can appear in the scope of quantifiers. Implicature-rich versions of (1), (3), and (5) are thus available — potentially usable by speakers and inferable by listeners just like any other semantic resolution for an underspecified form in context.

As we noted in the introduction, CFS draw a firm rhetorical distinction between their proposal and the Gricean approach to pragmatics. They state, "the goal of this paper is to challenge the neo-Gricean approach to SIs" (p. 2303), and, as we said, they later write that "the facts suggest that SIs are not pragmatic in nature but arise, instead, as a consequence of semantic or syntactic mechanisms" (p. 2316). The sense in which their account reflects this position is clear: to characterize implicatures, we need not consider the interactional setting or try to model the speaker and hearer. Rather, we can just describe a specific class of logical forms.

This position is tempered by CFS's pervasive appeals to pragmatic reasoning, however. The authors' specific examples are generally placed in contexts that support the target implicatures by ensuring that they are relevant, informative, and truthful. They concede that "aspects of the Gricean picture are sound and effective" (p. 2299). And, in summarizing their account, they make explicit the role that pragmatics must play in helping discourse participants to coordinate on the right logical forms:

> one can capture the correlation with various contextual considerations, under the standard assumption (discussed in the very beginning of this paper) that such considerations enter into the choice between competing representations (those that contain the operator and those that do not). (p. 2317)

The coordination problem that Grice sought to solve therefore remains, in the following form. First, in CFS's theory, the discourse participants must coordinate on the nature of the function *ALT*. Second, because the language permits but does not require silent, embedded *O* operators in many positions, the speaker's signal frequently underdetermines her intended message; a given surface form *U* might be consistent with logical forms that encode implicatures and those that don't, depending on where *O* appeared. Crucially, the speaker must rely on the listener to select the right one. Overall, then, implicature calculation now amounts to reasoning about which logical form was intended. How this coordination happens has not been a focus of grammar-driven accounts, but the above quotation suggests that communicative pressures like those Grice identified guide the process.

Summarizing so far, we have evidence from Chemla & Spector's (2011) experiments that some implicatures require, in some sense, local enrichment of embedded content via enhanced logical forms. Traditional Gricean accounts seem unable to capture such cases, but such accounts excel at characterizing how speakers and listeners coordinate on implicatures in simpler cases. CFS, in contrast, define a model in which local calculation is immediate, but they do not venture an account of how discourse participants coordinate on the right logical forms when more than one is allowed by the grammar. Stepping back, we see that both the Gricean and grammar-driven accounts clearly have something to contribute. We now turn to the task of developing a synthesis of the two approaches: a model that formally implements pragmatic reasoning over complex, compositionally defined logical forms and that is able to achieve the readings that seem to demand local enrichment. The technical details of the compositional model are different from CFS's, and

the technical details of the pragmatic account are different from Grice, but we hope that it combines the best aspects of both approaches.

4 A compositional lexical uncertainty model

We now present our mixed semantic–pragmatic model, which can be seen as a conceptual fusion of the Montagovian semantic perspective in Lewis (1970), the signaling systems of Lewis (1969), the probabilistic rational speech acts perspective of Frank & Goodman (2012) and Goodman & Stuhlmüller (2013), the iterated best response model of Jäger (2007, 2012) and Franke (2009), and the Bayesian view of Gricean reasoning developed by Russell (2012). Our Python implementation of the model is available from the website for this paper.

The model we implement here is a direct extension of the compositional lexical uncertainty model of Bergen et al. (2012) and Bergen et al. (2014) (see also Lassiter & Goodman 2013, 2015, for a closely related variant). This model defines production and interpretation as recursive processes in which speakers and listeners reason jointly about the state of world and the precise interpretation of lexical items in context. Our extension simply allows for greater diversity in the semantic lexicon and includes more complex aspects of semantic composition. Thus, in many ways, our central theoretical result is that Bergen et al.'s model predicts embedded implicatures in non-monotone and downward-monotone contexts if it is combined with a full theory of semantic composition.

The model's crucial feature is **lexical uncertainty**. In semantics, we like to imagine that word meanings are fixed across speakers and contexts, but in fact they are often idiosyncratic and adaptable (Clark & Clark 1979; Clark 1997; Lascarides & Copestake 1998; Glucksberg 2001; for an overview and general discussion, see Wilson & Carston 2007). Thus, in our model, discourse participants are not presumed to share a single, fixed lexicon mapping word forms to meanings. Rather, they consider many such lexica, and their communicative behavior, in both production and interpretation, is guided by their best attempts to synthesize the information from these varied sources (Giles et al. 1991). Thus, in the sentences of interest, the discourse participants might entertain multiple senses for an embedded *some*, including not only its 'at least' meaning but also the 'only some' meaning that corresponds to its enrichment by scalar implicature. This uncertainty carries through the compositional semantics to deliver embedded implicature readings. From this perspective, Chierchia et al.'s model is conceptually very close to lexical uncertainty, in that it requires reasoning about the logical form that a speaker intends to convey; a given token of *some* can take on multiple senses depending on the presence and nature of silent embedded operators in the logical form. Our extension of Bergen et al.'s model shows how this uncertainty guides pragmatic reasoning, and it furthermore shows that the uncertainty need not be fully resolved in order for robust pragmatic inferences to go through.

4.1 Grammar fragment

Table 1 gives the intensional fragment that we use throughout the remainder of this paper, both to explain how our pragmatic model works and to conduct our experimental analyses in section 6. It is our base lexicon, subject to refinement as part of pragmatic inference.

Syntax	Denotation of the lefthand side of the syntax rule
$\text{N} \rightarrow \textit{person}$	$\{\langle w, x\rangle : x \text{ is a person in } w\}$
$\text{N} \rightarrow \textit{shot}$	$\{\langle w, x\rangle : x \text{ is a shot in } w\}$
$\text{V}_\text{T} \rightarrow \textit{hit}$	$\{\langle w, x, y\rangle : x \text{ hit } y \text{ in } w\}$
$\text{V}_\text{I} \rightarrow \textit{scored}$	$\{\langle w, x\rangle : \exists y\ x \text{ hit } y \text{ in } w\}$
$\text{V}_\text{I} \rightarrow \textit{cheered}$	$\{\langle w, x\rangle : x \text{ cheered in } w\}$
$\text{D} \rightarrow \textit{some}$	$\{\langle w, X, Y\rangle : \{x : \langle w, x\rangle \in X\} \cap \{y : \langle w, y\rangle \in Y\} \neq \emptyset\}$
$\text{D} \rightarrow \textit{every}$	$\{\langle w, X, Y\rangle : \{x : \langle w, x\rangle \in X\} \subseteq \{y : \langle w, y\rangle \in Y\}\}$
$\text{D} \rightarrow \textit{no}$	$\{\langle w, X, Y\rangle : \{x : \langle w, x\rangle \in X\} \cap \{y : \langle w, y\rangle \in Y\} = \emptyset\}$
$\text{D} \rightarrow \textit{exactly one}$	$\{\langle w, X, Y\rangle : \lvert\{x : \langle w, x\rangle \in X\} \cap \{y : \langle w, y\rangle \in Y\}\rvert = 1\}$
$\text{NP} \rightarrow \textit{Player A}$	$\{\langle w, Y\rangle : \mathtt{a} \in \{x : \langle w, x\rangle \in Y\}\}$
$\text{NP} \rightarrow \textit{Player B}$	$\{\langle w, Y\rangle : \mathtt{b} \in \{x : \langle w, x\rangle \in Y\}\}$
$\text{NP} \rightarrow \textit{Player C}$	$\{\langle w, Y\rangle : \mathtt{c} \in \{x : \langle w, x\rangle \in Y\}\}$
$\text{NP} \rightarrow \text{D N}$	$\{\langle w, Y\rangle : \langle w, [\![\text{N}]\!], Y\rangle \in [\![\text{D}]\!]\}$
$\text{VP} \rightarrow \text{V}_\text{T}\ \text{NP}$	$\{\langle w, x\rangle : \{\langle w, y\rangle : \langle w, x, y\rangle \in [\![\text{V}_\text{T}]\!]\} \in [\![\text{NP}]\!]\}$
$\text{VP} \rightarrow \text{V}_\text{I}$	$[\![\text{V}_\text{I}]\!]$
$\text{S} \rightarrow \text{NP VP}$	$\{w : \langle w, [\![\text{VP}]\!]\rangle \in [\![\text{NP}]\!]\}$

Table 1: Interpreted grammar fragment. The left column defines a context-free grammar, and the right column gives its recursive interpretation in an intensional model $\langle D, W, [\![\cdot]\!]\rangle$, where D is a set of entities, W is a set of possible worlds, and $[\![\cdot]\!]$ is a semantic interpretation function. Notational conventions: $x, y \in D$, $w \in W$, and $X, Y \subseteq (W \times D)$.

The formal presentation is influenced by that of Muskens (1995): all of the denotations are sets, and the rules of semantic composition (the final four lines) combine them using operations that are formally akin to functional application. Our motivation for this less familiar presentation is that it makes it easy to define a uniform notion of refinement throughout the lexicon.

4.2 Refinement

The grammar in table 1 contains both lexical entries and rules of semantic combination. We assume that the rules are fixed. The lexical entries, on the other hand, are merely a starting point for linguistic communication — a set of somewhat negotiable conventions. You might assume that *couch* and *sofa* are synonymous, but if I say "It's a couch but not a sofa", you'll learn something about my lexical representations and perhaps adjust your own accordingly for the purposes of our interaction. If a speaker uses the phrase *synagogues and other churches*, then the listener can conclude that the speaker regards a synagogue as a kind of church, via the presuppositional nature of the phrase. Conversely, if the speaker says *church or synagogue*, the listener receives a weak signal that the speaker regards those two terms as disjoint, via the pressure for disjuncts to be exclusive (Hurford 1974). Chemla (2013) and Potts & Levy (2015) explicitly investigate such

listener implicatures and how they can be anticipated and potentially forestalled by speakers.

The 'lexical uncertainty' aspects of our model are designed to capture this variability. The core notion is that of lexical **refinement**, as defined in (11) following Bergen et al. (2014):

(11) a. Let φ be a set-denoting expression. R is a **refinement** of φ iff $R \neq \emptyset$ and $R \subseteq [\![\varphi]\!]$.

b. $\mathcal{R}_c(\varphi)$, the set of refinements for φ in context c, is constrained so that $[\![\varphi]\!] \in \mathcal{R}_c(\varphi)$ and $\mathcal{R}_c(\varphi) \subseteq \wp([\![\varphi]\!]) - \emptyset$

The full possible refinement space for a lexical item is the power set of its denotation minus the empty set. In a functional presentation of the interpreted fragment, this could instead be defined in terms of the subfunctions of a given denotation using a cross-categorical notion of entailment. With (11b), we allow that contexts can vary in how much of the full refinement space they utilize. They can be as small as the original denotation (in which case the uncertainty is eliminated), or as large as the full power set (minus the empty set).

The guiding idea is that, in interaction, pragmatic agents reason about possible refinements of their lexical items, with the base lexical meaning serving as a kind of anchor to which each word's interpretation is loosely tethered. Intuitively, one can imagine that part of what it means to be a responsible interlocutor is to make inferences, based on the speaker's behavior, not only about the world information she would like to convey, but also about the precise meanings she intends the words she is using to carry in the context of the interaction.

As we noted above, CFS's model embodies a kind of semantic uncertainty very similar to that considered here. For any given expression that one hears, the speaker might have in mind its literal content $[\![\varphi]\!]$ or one of the many enrichments available with $O_{ALT}(\varphi)$ for different choices of *ALT*. Similarly, we admit the trivial refinement $R = [\![\varphi]\!]$ as well as enrichments (technically, subsets) of it. The major technical difference lies in how these sets of denotations enter into the compositional semantics. For CFS, the alternatives all contribute to a single denotation, whereas our model keeps the alternatives separate during semantic composition, synthesizing them only for pragmatic inference. In terms of figure 1, we saw that CFS's theory uses O_{ALT} to create a single refined meaning for *some shot*, represented by the set of sets in the gray box ('some, not all shots'). Our theory of refinement could create one lexicon for every non-empty subset of the green box. So, in addition to considering 'some, not all, shots', we admit lexica that produce $[\![\textit{some shot}]\!] = \{\{a, b, c\}\}$ ('every shot'), lexica that produce $[\![\textit{some shot}]\!] = \{\{a, b, c\}, \{a\}\}$ (no obvious paraphrase), and so forth. These are all potential results of $O_{ALT}(\textit{some shot})$ for some choice of *ALT*, and our theory can be regarded as one that reasons in terms of all of these options.

4.3 Pragmatic reasoning

Our pragmatic model combines the logical grammar of section 4.1 with the lexical refinements of section 4.2. The basic ingredients are given in (12). We take as given a context c, an interpreted fragment $\langle \mathcal{G}, D, W, [\![\cdot]\!] \rangle$ as in table 1, with context free grammar $\mathcal{G}$, a domain of entities D, a set of worlds W, an interpretation function $[\![\cdot]\!]$ interpreting expressions of $\mathcal{G}$ in these domains, and a refinement function $\mathcal{R}_c(\varphi)$ that is defined for all lexical items in $\mathcal{G}$. For convenience, we assume that W is finite; this simplifies the definition of the probability measures but is not otherwise crucial.

(12) a. M is a subset of the proposition-denoting expressions generated by $\mathcal{G}$. It is augmented with a null message $\mathbf{0}$ such that $[\![\mathbf{0}]\!] = W$.

b. $\mathbf{L} = \left\{\mathcal{L} : \text{ for all } w \in W, \mathcal{L}(\mathbf{0}, w) = 1, \text{ and for all } m \in M, \{w : \mathcal{L}(m, w) = 1\} \in \mathcal{R}_c(m)\right\}$

c. $P : \wp(W) \mapsto [0, 1]$ is a prior probability distribution over sets of worlds. (For notational convenience, we abbreviate $P(\{w\})$ as $P(w)$.)

d. $C : M \mapsto \mathbb{R}$ is a cost function on messages. For lexical items, costs are specified. For a nonterminal node A with daughters $B_1 \ldots B_n$, $C(A) = \sum_{i=1}^{n} C(B_i)$.

e. $P_{\mathbf{L}} : \wp(\mathbf{L}) \mapsto [0, 1]$ is a prior probability distribution over sets of lexica. (For notational convenience, we abbreviate $P_{\mathbf{L}}(\{\mathcal{L}\})$ as $P_{\mathbf{L}}(\mathcal{L})$.)

In this paper, we do not bias the prior distribution over states P or the prior distribution over lexica $P_{\mathbf{L}}$ in any way, assuming them to be flat. Since we do not have experimental measurements for the priors, this seems like the safest option. (For techniques for measuring and manipulating state priors, see Frank & Goodman 2012 and Stiller et al. 2011.) Similarly, we do not explore different cost functions on non-null messages, assuming all costs to be zero.[2] Our cost functions play a role only in disfavoring the 'null message' $\mathbf{0}$, which is stipulated to be true in all worlds in all lexica.

In the context of our model, the set of messages M creates a space of alternative utterances that can drive complex pragmatic reasoning, as we will see in section 4.4. However, while these alternatives play a crucial role in capturing implicatures, they do not suffice for embedded ones. Thus, our focus is on the space of lexica defined by (12b) given a certain set of relevant alternative messages, as in (12a). Clause (12b) specifies all of the possible lexica $\mathcal{L}$ given the original interpretation function $[\![\cdot]\!]$ and $\mathcal{R}_c$. It is the space opened up by these constructs that allows us to predict where and how embedded implicatures will be perceived as salient. It should be noted in this context that our decision to refine only lexical items, as in (12b), is made only for simplicity. We could also allow arbitrary words and phrases to be refined, as CFS in effect do.

With this background in place, we now specify the core lexical uncertainty model. It consists of three inter-related agents, as defined in (13). The agents are defined in terms of the cost function C, the state prior P, and the lexica in $\mathbf{L}$. We assume throughout that m is any message in M, w is any state in W, and $\mathcal{L}$ is any lexicon in the set $\mathbf{L}$.[3]

(13) a. $l_0(w \mid m, \mathcal{L}) \propto \mathcal{L}(m, w)P(w)$

b. $s_1(m \mid w, \mathcal{L}) \propto \exp\left(\log l_0(w \mid m, \mathcal{L}) - C(m)\right)$

c. $L(w \mid m) \propto P(w) \sum_{\mathcal{L} \in \mathbf{L}} P_{\mathbf{L}}(\mathcal{L}) s_1(m \mid w, \mathcal{L})$

The first two agents, l_0 and s_1, are fixed-lexicon agents, and the final listener L reasons over all of the lexica in $\mathbf{L}$. The most basic agent is l_0. It defines a conditional distribution over worlds w given messages m. It does this by simply combining the truth conditions, given numerically as $\mathcal{L}(m, w)$, with the state prior. Where $\mathcal{L}(m, w) = 1$, the value is proportional to the state prior value

[2] The model is mathematically invariant to across-the-board additive transformations of message costs, so assuming all non-null messages to have zero cost loses no generality.

[3] $P(a \mid b) \propto F(a)$ is read 'the value $P(a \mid b)$ is proportional to the value $F(a)$'. The exact value of $P(a \mid b)$ can always be obtained by dividing $F(a)$ by the normalizing constant $Z = \sum_{a'} F(a')$ so long as this sum is finite, which is guaranteed to be the case in the class of models defined in (12).

$P(w)$; where $\mathcal{L}(m, w) = 0$, the value is 0. So this is just the semantics turned into a probability distribution for the sake of decision making; the intuitive idea is that the agent hears m and estimates the relative likelihood of worlds on that basis.

The speaker agent s_1 is already a pragmatic agent, in the sense that it reasons not about the lexicon directly but rather about how the listener will reason about the lexicon. The speaker observes a state w and chooses messages on that basis. The logarithm and exponentiation in this definition allow us to include real-valued costs; where the costs are all 0, it reduces to $s_1(m \mid w) \propto l_0(w \mid m)$, by the identity $x = \exp(\log(x))$.[4]

Some comment is in order regarding the role of the null message in the model. Technically, $\mathbf{0}$ allows us to explore the full space of refinements for messages while guaranteeing that, for every possible speaker's observed state w, there is some compatible message m such that $\mathcal{L}(m, w) = 1$. Without this, the speaker distribution $s_1(m \mid w, \mathcal{L})$ would not be defined. There are a few alternative methods for addressing this technical issue. Bergen et al. (2012) admit only lexica in which the speaker has at least one true message for every state; Bergen et al. (2014) briefly consider giving false states tiny positive probability; and Jäger (2012) defines a belief-revision step to handle comparable situations in the context of the iterated best-response model. The null-message approach has qualitatively similar behavior to these other approaches, and we favor it here because it is technically simpler to implement.[5] We set $C(\mathbf{0}) = 5$ throughout the paper, but changing this value does not change our qualitative predictions. (See also Appendix A.)

Our pragmatic listener is defined in (13c). This agent resembles the literal listener l_0, but it sums over all of the inferences defined by the lexicon-specific agents s_1 and l_0. It additionally incorporates the state prior, as l_0 does, and the prior over lexica. This is the agent that we use to characterize listener inferences and define our predictions about our experimental findings.

We have presented the compositional lexical uncertainty model in its simplest form, but we have gone beyond Bergen et al. in three respects. We give a more complete treatment of semantic composition, we allow uncertainty in the denotations of lexical items of a wider range of semantic types, and we entertain the possibility of restrictions on the set of possible refinements. However, many other elaborations of models in this space are possible (Goodman & Lassiter 2015; Smith et al. 2013; Kao et al. 2014b; Potts & Levy 2015). Two particular elaborations are highly salient given the prior literature. First, one could allow further iteration beyond L, defining speaker and listener agents analogously to their fixed-lexicon counterparts. This can amplify existing pragmatic inferences and create new ones (Bergen et al. 2014; Vogel et al. 2014; Potts & Levy 2015). Second, one can include a real-valued temperature parameter λ in the speaker agents to control how greedily they try to extract information from the agent they are reasoning about, with higher λ values leading to more aggressive inferential strategies (Sutton & Barto 1998). This too can radically reshape the agents' behavior. In appendix A, we explore the consequences of these elaborations

[4] We could equivalently define an alternative cost function C' ranging over $[0, \infty)$ such that $C'(m) = e^{C(m)}$, and then replace (13b) with $s_1(m \mid w, \mathcal{L}) \propto l_0(w \mid m, \mathcal{L})C'(m)$.

[5] A closely related alternative, technically more complex but perhaps more pretheoretically transparent, would be to posit a collection of "null" messages, one for each speaker's observed state, each admitting only that state, and each having a considerably higher cost than all the non-null messages. This alternative has the interpretation that the null messages constitute the collection of more precise but much more prolix utterances the speaker might have used to describe her observation state. The behavior of this alternative approach would be qualitatively the same as ours: the specialization of each null message for a unique world state would strengthen its appeal for s_1, but its high cost would countervail that appeal.

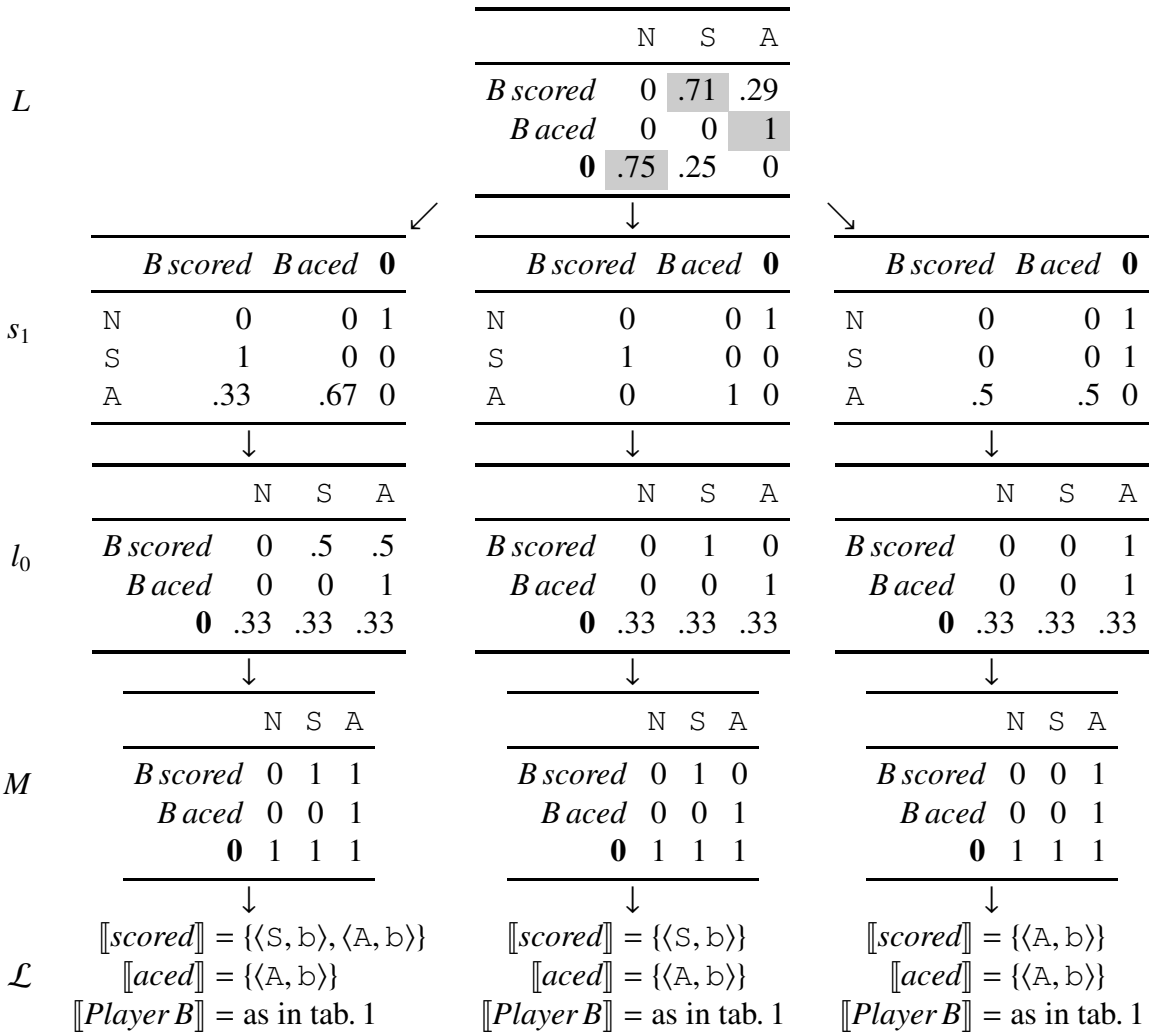

Figure 2: Simple scalar inference. We assume a flat prior over states and lexica. $C(\mathbf{0}) = 5$, and $C(m) = 0$ for the other messages. The uncertainty listener L infers that the general term *scored* excludes its specific counterpart *aced* in this context.

for modeling the pattern of data we observed.

4.4 Illustrations

Our first illustration, given in figure 2, is designed solely to reveal details about how the agents interact to produce enriched construals. (This first illustration is isomorphic to the example covered in section 4.4 of Bergen et al. 2014.) We assume that the domain consists of just one entity, b, and that the only intensional distinction of interest is whether b scored none of his shots (world N), some but not all of his shots (S), or all of his shots (A). The action is in the relationship between the two predicates *scored* and *aced*: we define $[\![scored]\!] = \{\langle S, b\rangle, \langle A, b\rangle\}$ and $[\![aced]\!] = \{\langle A, b\rangle\}$. Thus, *aced* strictly entails *scored*, creating the potential for an SI.

To keep the example compact, we let $\mathcal{R}_c(\textit{Player B}) = \{[\![\textit{Player B}]\!]\}$. Since *aced* already denotes a singleton set, it has no space for further refinement. However, *scored* has two further refinements. This gives rise to the three lexica in the bottom row of figure 2. Using the fixed rules of semantic composition, these lexica determine the messages *Player B scored* and *Player B aced*. The literal listener l_0 turns the denotations of these messages into conditional distributions over states given messages. The prior over states is flat in this example, so this calculation just evenly divides the probability mass over the true states. The pragmatic speaker responds to this agent. Finally, our uncertainty listener sums over these three speakers. This listener achieves an SI in the following nuanced, probabilistic sense (Russell 2012: §2). Hearing *Player B scored* leads this listener to assume that the most probable state is `S`. The probability is not 1, so uncertainty remains. However, if this listener is compelled to make a categorical decision about the intended meaning of the utterance, he will choose this enriched construal, and he will rightfully feel deceived if the world state turns out to be `A` or (worse) `N` instead. In this way, the model characterizes the uncertainty surrounding implicature inferences (Hirschberg 1985) and the ways in which this uncertainty relates to decision making.

Lexical uncertainty is not required to achieve this result. If we allow no meanings to be refined, then we deal with the singleton set of lexica containing only the leftmost lexicon. In this small space, the model shares deep affinities with the Bayesian model of Gricean reasoning given by Russell (2012); it is effectively equivalent to the rational speech act model of Frank & Goodman (2012) (potentially with small differences relating to how the prior over states is incorporated); and it can be seen as a more thoroughly probabilistic version of the iterated best response model (Franke 2009; Jäger 2007, 2012). Nonetheless, the example illuminates how the lexical uncertainty model works. As the downward arrows indicate, it is useful to start conceptually from L. This agent effectively reasons in Gricean terms about three separate lexica; the alternation from speaker to listener and down to the lexicon mirrors the nested belief structure of Grice's original definition of implicature (sketched at the start of section 2).

Even though we assume an even prior over lexica, useful biases emerge because the space of lexica is structured: there are no lexica in which *aced* is consistent with `S`, but there are two in which *scored* is. This bias carries through the computation to create a strong final bias for the implicature inference. For further discussion of this important point, we refer to Bergen et al. 2014, where it is shown to be essential to generating implicatures based on the principle that marked forms signal marked meanings and unmarked forms signal unmarked meanings (McCawley 1978; Horn 1984; Blutner 1998; Levinson 2000).

The lexical uncertainty aspects of the model are a rich source of implicatures, and they are the key to achieving local implicatures of the sort reviewed in section 2 above. However, as the fixed lexicon versions of the model make clear, the recursive nature of the agents suffices for many kinds of enrichment assuming the space of alternative messages M is chosen properly. Even with a single lexicon, we have a listener reasoning about a speaker reasoning about the literal interpretive semantics, which creates forces for removing semantic overlap among the alternative messages. One powerful illustration of this comes from Sauerland (2001, 2004), who studies the implicatures of sentences like *Player A hit some of his shots or cheered*, in which the weak scalar term *some of his shots* is nested inside the weak connective *or*. The guiding intuition is that the sentence is most prominently construed as entailing that Player A did not make all of his shots and that Player A did not both make shots and cheer. Sauerland's insight is that these entailments are within reach of traditional neo-Gricean reasoning as long as the available alternative messages that

	•	C	S_1	S_2	S_1C	S_2C	S_1S_2	S_1S_2C
Player A cheered	0	.43	0	0	.23	.23	0	.10
Player A hit every shot	0	0	0	0	0	0	.72	.28
Player A hit some shot	0	0	.33	.33	.09	.09	.10	.04
Player A hit some shot or cheered	0	.15	.28	.28	.08	.08	.09	.03
Player A hit some shot and cheered	0	0	0	0	.41	.41	0	.17
Player A hit every shot or cheered	0	.34	0	0	.19	.19	.20	.08
Player A hit every shot and cheered	0	0	0	0	0	0	0	1
0	1	0	0	0	0	0	0	0

Table 2: Inferences from nested scalar terms arising from competition among messages alone. (Introducing lexical uncertainty into the model only strengthens the basic patterns seen here.)

the speaker might have used is comprehensive in that it fully crosses the alternatives for *some* with the alternatives for *or*.

As table 2 shows, our model suffices to achieve this even with a fixed lexicon. For simplicity, we assume there are just two shots in the domain. Columns indicate the truth values of individual predicates: in S_1, Player A made the first shot, missed the second, and didn't cheer; in S_1S_2, Player A made every shot but didn't cheer; in C, Player A made no shots but cheered; in •, Player A made no shots and didn't cheer; and so forth. The crucial analytic step is to define the set of messages M so that it covers the space that Sauerland described. This suffices to capture the desired inferences in the probabilistic sense that our model provides: given *Player A hit some shot or cheered*, our pragmatic listener (13c) places most of its probability mass on worlds in which Player A only cheered (C) or made only some shots and did not cheer (S_1, S_2). We also see the expected scalar inferences from *some* and *or* when they appear independently: *Player A hit some shot and cheered* leads the listener away from states where both shots were made, and *Player A hit every shot or cheered* leads the listener away from the world verifying both conjuncts, S_1S_2C.

We obtained the results of table 2 using a uniform prior over states, but similar qualitative patterns would hold using a different prior specification. Likewise, allowing lexical refinements, as in the full version of our model, strengthens the relevant inferences without changing the qualitative pattern seen in table 2. For brevity we do not show this result, but readers are encouraged to try the simulations for themselves, using the code provided with this paper.

Let's now look at a larger and more complex scenario, one in which lexical uncertainty interacts with message competition to help reveal the potential of this model to capture embedded implicatures in ways that a fixed-lexicon version of the model cannot. In this scenario, there are two players. We resume our convention of referring to worlds using sequences like NN ('neither player scored'). The lexical items are *Player A*, *Player B*, *some*, *every*, *no*, *scored*, and *aced*. To start, we assume that, for all lexical items φ, $\mathcal{R}_c(\varphi) = \wp(\llbracket\varphi\rrbracket) - \emptyset$. This creates an enormous space of lexica, and allows the full range of possible interactions between the refinements.

The listener inferences are summarized in table 3. For the most part, they seem aligned with the general view in the literature about how scalar terms interact in contexts like this. For instance, we predict that a proper name P will take on the exhaustified sense *only P*, as we would expect given the salience of *every*. In turn, *some* is interpreted as non-specific in virtue of the salience of

the two names, and it also leads to an SI due to the salience of *every*. Perhaps the most striking outcome is that the scalar inference from *scored* to not-aced remains in evidence not just with the proper names but also in the scope of the quantified subjects: the best-guess inference for *every player scored* is SS. These effects derive from interacting lexical uncertainty between the subjects and predicates.

Table 3 reveals some drawbacks to unfettered exploration of refinements, however. First, we might expect hearing *some player scored* to lead the listener to assume that the state was either NS or SN, corresponding to enrichment of both the subject ('not all players') and the predicate ('merely scored'). The current model does not achieve this. In addition, the row for *no player scored* is unintuitive. The best inference is NN, which is in line with the literal semantics, but it is striking that the states NS and SN have some positive probability. This arises because of interacting lexical uncertainty: there are lexica in the space in which *scored* is refined to exclude one of the players. In that case, the negative universal turns out to be true. Only a few lexica support this interaction, ensuring that it cannot become dominant, but it still seems worrisome.

This worry is a touchstone for revisiting an assumption of the model underlying table 3: that the lexical items can be refined in completely arbitrary ways. We take it to be one of the major lessons of neo-Gricean approaches that alternatives are contextually and lexically constrained. CFS's treatment of *ALT* reflects this lesson, as do our own sets of alternative messages M. Our handling of refinement allows us to incorporate such insights at the level of lexical uncertainty as well. This is not part of the neo-Gricean perspective as normally construed, but it's a natural step in the context of our model. Thus, it is worth seeing whether we can improve the picture in table 3 by encoding lexical scales in our grammar fragment.

We implement lexical scales in our model by constraining the refinement sets for several lexical items, as follows:[6]

(14) a. $\mathcal{R}_c(\textit{Player A}) = \{[\![\textit{Player A}]\!], [\![\textit{only Player A}]\!]\}$

b. $\mathcal{R}_c(\textit{Player B}) = \{[\![\textit{Player B}]\!], [\![\textit{only Player B}]\!]\}$

c. $\mathcal{R}_c(\textit{some}) = \{[\![\textit{some}]\!], [\![\textit{some and not all}]\!]\}$

d. $\mathcal{R}_c(\textit{no}) = \{[\![\textit{no}]\!]\}$

e. $\mathcal{R}_c(\textit{scored}) = \{[\![\textit{scored}]\!], [\![\textit{scored and didn't ace}]\!]\}$

f. $\mathcal{R}_c(\textit{aced}) = \{[\![\textit{aced}]\!]\}$

The results of working in this more constrained, neo-Gricean refinement space are given in table 4. The picture is mostly unchanged, except we now also achieve the target enrichment for *some player scored*, and the messiness surrounding *no player scored* is fully addressed. The one remaining potential concern about table 4 is that it predicts rather aggressive pragmatic enrichment of the scalar term in the scope of the negative quantifier. As we noted in section 2, it has long been assumed that weak scalar items in such environments fail to give rise to upper-bounding implicatures. Chemla & Spector (2011) address this question empirically, finding in their experiment low but non-negligible rates of local enrichment in negative environments. We too treat this as an empirical question; in section 6, we present evidence that local enrichments of this sort are indeed salient possibilities for humans.

[6] We define $[\![\textit{only Player A}]\!] = \{\langle w, Y\rangle : \{\mathtt{a}\} = \{x : \langle w, x\rangle \in Y\}\}$, and similarly for $[\![\textit{only Player B}]\!]$, not as a claim about natural language *only*, but rather just for the sake of the simulation.

	NN	NS	NA	SN	SS	SA	AN	AS	AA
Player A scored	0	0	0	.24	.19	.16	.18	.16	.07
Player A aced	0	0	0	0	0	0	.36	.30	.34
Player B scored	0	.24	.18	0	.19	.16	0	.16	.07
Player B aced	0	0	.36	0	0	.30	0	0	.34
some player scored	0	.14	.11	.14	.17	.14	.11	.14	.05
some player aced	0	0	.22	0	0	.19	.22	.19	.18
every player scored	0	0	0	0	.31	.27	0	.27	.14
every player aced	0	0	0	0	0	0	0	0	1
no player scored	.31	.14	.12	.14	.06	.05	.12	.05	.01
no player aced	.18	.19	.08	.19	.14	.06	.08	.06	0
0	.01	.01	.32	.01	.01	.15	.32	.15	0

Table 3: Enrichment in the largest space of refinements supported by this lexicon.

	NN	NS	NA	SN	SS	SA	AN	AS	AA
Player A scored	0	0	0	.45	.11	.22	.15	.05	.02
Player A aced	0	0	0	0	0	0	.42	.36	.22
Player B scored	0	.45	.15	0	.11	.05	0	.22	.02
Player B aced	0	0	.42	0	0	.36	0	0	.22
some player scored	0	.25	.09	.25	.06	.12	.09	.12	.01
some player aced	0	0	.24	0	0	.21	.24	.21	.11
every player scored	0	0	0	0	.61	.16	0	.16	.07
every player aced	0	0	0	0	0	0	0	0	1
no player scored	.61	0	.16	0	0	0	.16	0	.06
no player aced	.19	.17	.10	.17	.13	.07	.10	.07	0
0	.15	.13	.13	.13	.10	.09	.13	.09	.05

Table 4: Enrichment using the lexically-driven (neo-Gricean) refinement sets in (14).

5 Prior experimental work

The illustrative examples in the previous section begin to show that our compositional lexical uncertainty model naturally generates local enrichments. Thus, the question of whether listeners actually make such inferences is critical in judging the suitability of this model as a description of human reasoning. The present section reviews the prior literature in this area.

The pioneering paper is Geurts & Pouscoulous 2009. Their experiments 3 and 4 asked participants to provide truth-value judgments for sentences paired with abstract visual scenes consisting of shapes connected by lines. The target sentences included weak scalar terms in upward, downward, and non-monotone contexts, such as *exactly two of the squares are connected with some of the circles*, comparable in relevant respects to the examples reviewed in section 2 above. Geurts & Pouscoulous found only negligible rates of inferences consistent with local enrichment. These findings stimulated a number of responses commenting on the prevalence of local enrichment and its theoretical import (Ippolito 2010; Sauerland 2010). The two responses that are most

relevant for our purposes are those of Clifton & Dube (2010) and Chemla & Spector (2011).

Clifton & Dube (2010) argue that the experimental setting used by Geurts & Pouscoulous was prone to understating the rate of implicatures, and they sought to address this issue with a different experimental method. In their experiment, one trial consisted of presenting the participant with a sentence together with a set of visual scenes. The participant was instructed to choose the scene or scenes, if any, that he or she considered "best described by the sentence". They found that participants tended to chose the scene consistent with local enrichment. This method is a natural choice given a pragmatic model like ours, since it places participants in a role comparable to that of a listener agent. The particulars of the experimental method were criticized by Geurts & van Tiel (2013: §5.1) and van Tiel (2014), however, on the grounds that, for the examples involving monotone quantifiers, the inferences are better explained in terms of the typicality effects of the quantifiers involved (see also Degen & Tanenhaus 2015). Roughly speaking, the claim is that the typicality structure of *some A are B* favors situations in which just shy of half the A's are B's, and experimental designs (like Clifton & Dube's) that allow participants to express extra-truth-conditional preferences will be sensitive to this typicality structure. While we think that typicality is an important component of many implicatures and thus should ultimately be derived from a complete pragmatic model rather than considered a separate, overlaid factor,[7] we also see value in trying to neutralize its effects for purposes of studying local enrichment.

Chemla & Spector (2011) followed Geurts & Pouscoulous (2009) in asking participants to interpret quantified sentences in abstract geometric scenes, but they sought to simplify those scenes (see Geurts & van Tiel 2013: 31 for criticisms of this presumption), and they allowed subjects to provide graded truth-value judgments on a scale between 'Yes' and 'No'. The results were consistent with very high rates of local enrichment in upward and non-monotone environments, and even yielded suggestive evidence for local enrichment in downward monotone environments. These findings stand in stark contrast to those of Geurts & Pouscoulous (2009).

However, there are at least three features of Chemla and Spector's experimental design that might have exaggerated the rates of judgments consistent with local enrichment (Geurts & van Tiel 2013). First, the graded response categories mean that, for the monotone cases, typicality effects might have played a role. Second, the visual scenes were wheel-like displays in which lines extend from the vertex to the perimeter. There are potentially many ways this state can be drawn. Some might be more iconic than others, and some might create spurious patterns and salience contrasts that could affect linguistic inference in unmeasured ways. Third, Chemla & Spector used a within-subjects design: the individual participants judged every sentence in every context. Participants could thus have drawn comparisons across different conditions, creating opportunities for them to register comparative judgments involving the experimental contexts themselves, rather than relying solely on their linguistic intuitions.

We draw three major lessons from the above studies and debates. First, we should seek out simple, naturalistic stimuli. Previous experiments in this area have used abstract displays. Together with the inevitable complexity of the sentences involved, this choice seems likely to put cognitive demands on participants in ways that could affect the stability and reliability of the responses. Second, scalar response categories might encourage typicality inferences that could cloud

[7] Levinson's (2000) I-implicatures involve inferences from a general term or statement to one of its salient or prototypical subkinds. In the context of a generalized theory of scalar (partial-order) inference like that of Hirschberg (1985), this can be seen as a scalar inference guided by prior expectations.

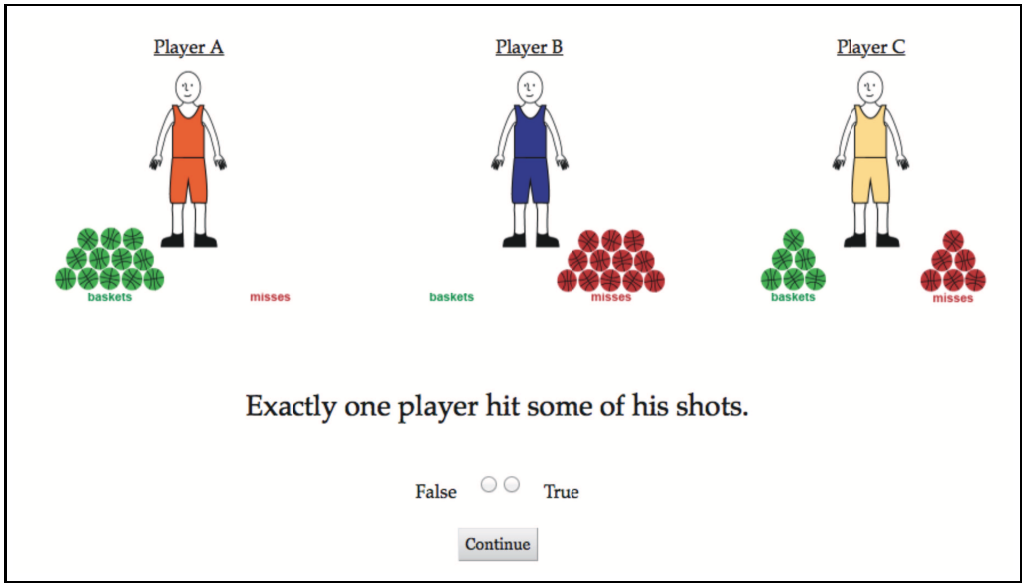

Figure 3: Experiment display.

the implicature picture; this might be a concern only for monotone environments, but we can hope to avoid the issue by restricting to just truth-value responses. Third, to the greatest extent possible, we should seek a design that supports analyses in which we can marginalize out the idiosyncrasies of particular displays, to avoid artifacts of salience or contrast that could stimulate responses that are consistent with implicature calculation without requiring such calculation.

6 Experiment: Scalars under quantifiers

We now present our main experiment involving *some* in quantified environments. We told participants that they were helping to train an automated sportscasting system and asked them to provide truth-value judgments about sentences in the context of displays like figure 3. This cover story was designed to ensure that implicatures are relevant, that is, worth calculating where available (Chemla & Spector 2011: §3.1; Clifton & Dube 2010). Our goal was to better understand the extent to which certain pragmatic inferences are available, so we sought out a scenario that would be maximally favorable to them. (For studies aimed at understanding the prevalence of implicatures, see Paris 1973; Hendriks et al. 2009; Degen 2015.)

6.1 Methods

6.1.1 Participants

The experiment had 800 participants, all recruited with Amazon's Mechanical Turk. No participants or responses were excluded.

6.1.2 Materials

We generated displays like those in figure 3. In each display, each of the three players, A, B, and C, has taken 12 basketball shots (a number small enough for visual display but outside of the subitizing range and thus less likely to introduce competitions from cardinal determiners like *three shots*; Degen & Tanenhaus 2015). The shots were divided into two piles, labeled 'baskets' (green) and 'misses' (red). For our target items, the player either made all 12 baskets (Player A in figure 3), missed all 12 baskets (Player B), or made 6 and missed 6 (Player C). The colors of the players' clothes were set randomly from a palette of 14 colors.

The target sentences describing the displays were defined as follows:

(15) $\left\{\begin{array}{l}\text{Every}\\ \text{Exactly one}\\ \text{No}\end{array}\right\}$ player hit $\left\{\begin{array}{l}\text{all}\\ \text{none}\\ \text{some}\end{array}\right\}$ of his shots.

Following previous studies, we put a bound pronoun in the embedded quantifier to try to ensure that the subject took scope over the object. The partitive forms seem likely to further encourage implicature calculation (Reed 1991; Grodner et al. 2010; Degen 2015). We chose the verb *hit* over the slightly less marked verb *make* to try to avoid the sense of 'make' as in 'take' (consistent with missing).

For the target items, there were ten different conditions, corresponding to the worlds in (16), in the notation we've been using to identify possible worlds.

(16) {NNN, NNS, NNA, NSS, NSA, NAA, SSS, SSA, SAA, AAA}

This is a subset of the full cross-product of the three outcomes N, S, and A in which player i always did at least as well as player $i+1$, going left to right. Our target sentences were all quantified, so we don't care about the outcome for any single player, meaning that we don't distinguish, e.g., NNS from NSN, allowing us to work with this smaller set of conditions. In the experiment, the 'order' of each world was randomized, so that, e.g., NSA appeared visually in each of its three orders approximately the same number of times. This randomization allows us to control for preferences in visual processing that might naturally make one position or linear ordering of player outcomes salient in unanticipated ways.

6.1.3 Procedure

After reading our consent form, participants were given the following cover story about "a basketball free-throw shooting competition between 3 players":

(17) We are trying to train an automated sportscasting system to generate color commentary on simple competitions. We'd like you to make judgments about the comments it generates. We'll use these ratings to train our system further.

After reading this cover story and some instructions, participants were presented with three training items, designed to ensure that participants understood the cover story, displays, and sentences. They then judged 32 sentences, divided into 9 target sentences and 23 fillers. The design was between-subjects: no experimental participant judged the same sentence twice. The order of presentation of the items was randomized.

Each sentence received a total of 800 responses. For the target sentences, each sentence–world pair received between 58 and 103 responses (mean 80); this variation resulted from randomization in the assignment of worlds to sentences.

Target sentences were presented below displays. Participants were asked to evaluate sentences as either true or false. In this sense, our participants acted as listeners who got to observe the speaker's state and assess whether the speaker accurately described that state with her utterance. We also conducted a variant of the experiment in which participants gave responses on a seven-point Likert scale ranging from 'Bad description' to 'Good description', to see whether this would reveal information about the quality of the report. These two versions of the experiment led to qualitatively identical outcomes. Appendix B reviews the details of the scalar-response version.

All the materials and response data for the experiment are available at the website for this paper.

6.2 Results

Figure 4 summarizes the responses by target sentence and the world in which it was evaluated. Overall, participants made judgments that accurately reflected whether sentences were true or false; accuracy was especially clear for the sentences in the first two columns, which do not admit pragmatic enrichment. For these cases, the responses were essentially categorical. This pattern suggests that our method is appropriate for measuring participants' interpretations.[8]

We now turn to the critical conditions, reporting significance levels for key theoretical comparisons based on the nonparametric Mann–Whitney U test. Responses for '*every…some*' (upper right) were consistent with the hypothesis that *some* is locally enriched in this condition. In particular, this sentence received the greatest percentage of 'True' responses in the SSS world. As we reviewed in section 2, in order to count as a complete report in this world, this sentence requires either local enrichment or a Gricean calculation with auxiliary premises. Worlds SSA and SAA received the next highest percentages of 'True' responses (lower than SSS, $p = 0.09$ and $p = 0.04$, respectively). Of all the literally true worlds for this condition, AAA received the lowest percentage of 'True' responses (lower than SSA and SAA; both at $p < 0.01$). Only a simple Gricean calculation is required to account for the higher rate of 'True' for SSA and SAA compared with AAA: in the latter world, the salient alternative *every player hit all of his shots* is a more complete description.

Nevertheless, '*every…some*' is not a strong test of the presence of local readings, since the entailment relations between the readings introduce some indeterminacy into the analysis. In particular, since the local enrichment entails the literal reading, we can't be sure whether the 'True' responses for SSS derive entirely from the availability of a local enrichment: a literal construal would suffice to make the sentence true. Furthermore, as discussed in section 2, '*every…some*' is of limited utility in distinguishing theoretical proposals anyway. It is the '*exactly one…some*' sentence that allows us to probe most confidently for local readings.

The response pattern for the critical item '*exactly one…some*' is given in the middle right of

[8]The only exception to this general pattern is the sentence *No player hit none of his shots* (bottom middle). The percentage of 'True' responses is lower than normal in all its true conditions and relatively high for NNN, where it is false on its literal construal. We hypothesize that this pattern reflects a negative concord construal, on which the embedded term is interpreted as equivalent to *any of his shots*, creating a meaning that is true only in NNN. Negative concord of this sort is productive in many dialects of English and understandable in all or nearly all of them. This likely created uncertainty about the intended meaning of the sentence, leading participants to disfavor it in general.

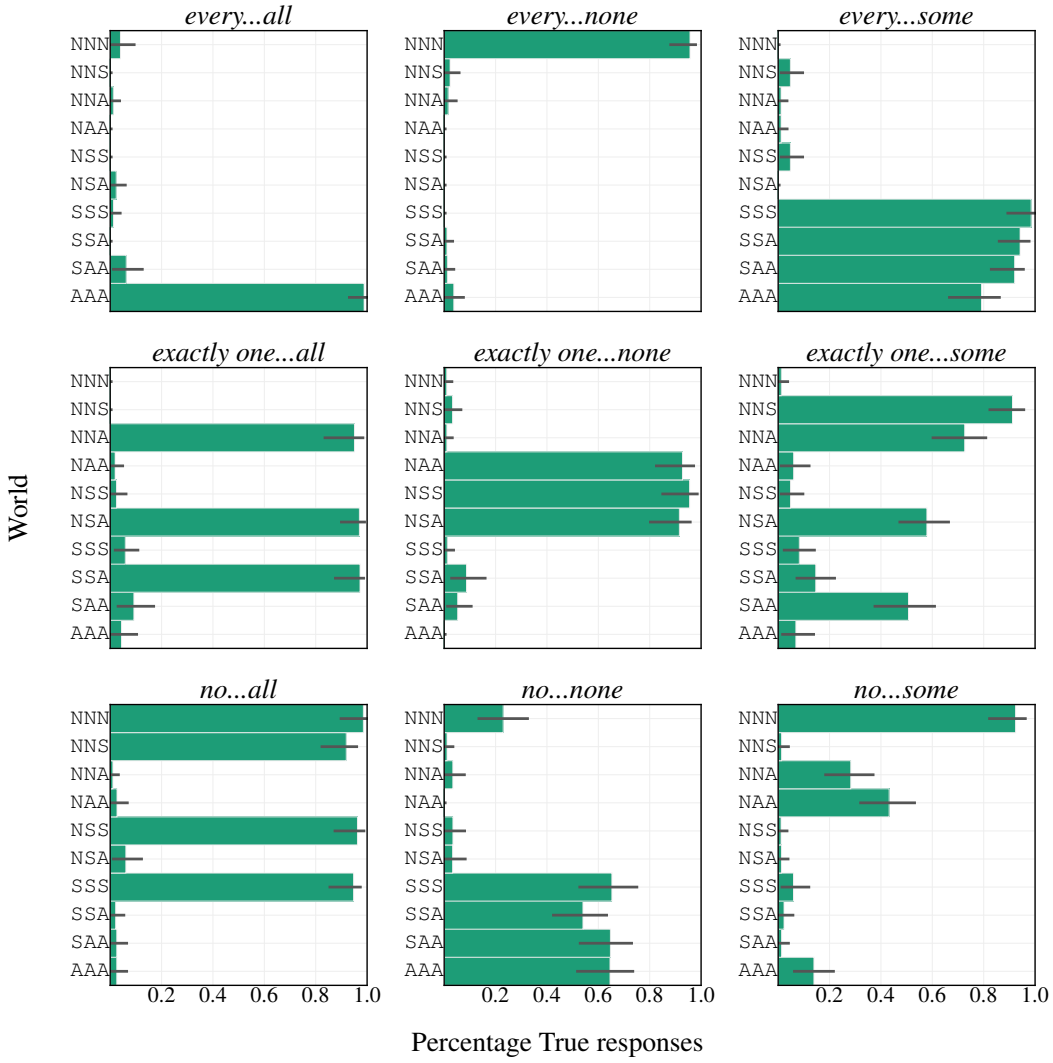

Figure 4: Mean truth-value judgment responses by sentence with bootstrapped 95% confidence intervals.

figure 4. The highest percentage of 'True' responses is for the NNS condition, where the sentence is true under its literal and local enrichment construals. However, it was also frequently judged true in the NSA and SAA worlds (both higher at $p < 0.001$ than in SSA, the world yielding the highest rating among those in which the sentence is false both literally and under all possible enrichments). For NSA and SAA, the sentence is true only with local enrichment (because two players hit at least some of their baskets in these worlds, ruling out the literal construal). We note also that its more strictly truth-conditional interpretation seems to be salient as well, as it was generally perceived to be true in the NNA condition.

Finally, the pattern for '*no…some*' also suggests a non-trivial amount of local enrichment: though NNN produced the highest rate of 'True' responses, indicating a preference for a literal construal, the 'True' rates for NNA, NAA, and AAA are consistently higher than for the most favored false worlds, NNS and NSA; all pairwise significance tests for the cross-product of {NNS, NSA} and {AAA, NNA, NAA} are significant at $p = 0.002$. These are the worlds in which no player hit only

some of his shots, the local enrichment. This finding seems consistent with the low but non-negligible rates of local enrichment that Chemla & Spector (2011: §4.4.4) report for this quantifier pair. One qualification we should add here is that our sentence is arguably somewhat unnatural in that it places *some*, a positive polarity item (Baker 1970; Israel 1996), in the scope of a negative quantifier. The binding relation between the subject and the pronoun *his* in the embedded phrase should force a surface-scope reading, but we can't rule out the possibility that participants might have found an inverse-scope construal ('some shots are such that no player hit them') that took the scalar term out of the scope of the negation. Alternatively, the marked nature of *some* in this position might have encouraged implicit prosodic focus, which would also likely increase the 'only some' construals.

We conclude from these responses that local enrichment is possible even in non-monotone environments, and that local enrichment might be available in downward-monotone environments as well. However, our concern is not only whether such readings are possible or impossible, but rather how accurately we can predict their availability on the basis of contextual and world knowledge. We turn now to the task of assessing the ability of the model presented in section 4 to match both the quantitative and qualitative patterns in our experimental data.

6.3 Model assessment

The pattern of data we observed is sufficiently precise and detailed that extracting its full theoretical value requires more than arbitrary statistical tests of simple null hypotheses — e.g., the null hypothesis that in the '*exactly one…some*' condition, ratings are the same for the worlds admitted by local enrichment as for those excluded under both global and locally-enriched interpretations. This and other such null hypotheses can be rejected with high confidence. Instead, to characterize the patterns of inference that give rise to the observed data, we use a model-comparison approach. In particular, we evaluate four related models that each embody different characterizations of linguistic meaning. By comparing these models, we can gain insights into the aspects of each that contribute to particular patterns of predictions.

Our assumption in this comparison is that our models provide a description of aggregate human behavior across individuals. In this sense, they are posed at Marr's (1982) 'computational theory' level. They instantiate claims about the task that our participants are attempting to perform and the assumptions that they use in performing it, but they are agnostic about the particular processes ('algorithms', in Marr's terminology) by which individuals perform it. In particular, the averaged binary responses that we take as our modeling target could come about via a number of routes. For example, individuals could be computing or approximating the full computations that we describe here and then stochastically making binary choices based on their estimates of the underlying probability distribution. Alternatively, they could also be pursuing any number of heuristic, approximate strategies that — when aggregated across individuals and trials — could yield a stable probability estimate. We remain agnostic about this issue here, but we note that a growing literature explores these different hypotheses linking computational-level models to psychological processes (e.g., Bonawitz et al. 2014; Griffiths et al. To appear; Sanborn et al. 2010; Vul et al. 2014).

For all the models, we take as given the literal semantics described in table 1, as well as the following features of the context:

(18) a. $D = \{\texttt{a}, \texttt{b}, \texttt{c}\}$

b. W = the set in (16)

c. $M = \left\{ Q(player)(hit(S(shot))) : \begin{array}{l} Q \in \{exactly\ one, every, no\}, \\ S \in \{every, no, some\} \end{array} \right\} \cup \{\mathbf{0}\}$

d. $C(\mathbf{0}) = 5$; $C(m) = 0$ for all $m \in M - \{\mathbf{0}\}$

e. Flat state prior: $P(w) = P(w')$ for all $w, w' \in W$

f. Flat lexicon prior: $P_{\mathbf{L}}(\mathcal{L}) = P_{\mathbf{L}}(\mathcal{L}')$ for all $\mathcal{L}, \mathcal{L}' \in \mathbf{L}$

The domain D and worlds W come directly from our human experiment. Similarly, the set of messages M corresponds to (15), with some adjustments to keep the logical grammar simple. We stipulate flat priors and even costs (other than the null message). As noted in section 4, we do not have empirical estimates for these values; though better fits to the human data can be achieved by adding assumptions about them, this risks overfitting to the particular data we have and thus overstating the true accuracy of the models. The value $C(\mathbf{0}) = 5$ was chosen arbitrarily; appendix A explores a wide range of values for it.

The models we consider are defined as follows:

(19) a. **Literal semantics**: the predicted values are the output of l_0, as in (13a), run on the messages defined in (18c).

b. **Fixed-lexicon pragmatics**: the predicted values are the output of the uncertainty listener (13c), but all the lexical items have only themselves as refinements, so that the reasoning is entirely in terms of the base lexicon in table 1.

c. **Unconstrained refinement**: the inferences of the uncertainty listener (13c) with $\mathcal{R}_c(some) = \wp([\![some]\!]) - \emptyset$

d. **Neo-Gricean refinement**: as in 'Unconstrained refinement', but with $\mathcal{R}_c(some) = \{[\![some]\!], [\![some\ and\ not\ all]\!]\}$, as in (14) of section 4.4, to extend neo-Gricean insights about alternatives into the lexical uncertainty aspects of our model.

These models represent a broad range of approaches to linguistic meaning. The first neglects pragmatics entirely (the model includes a contextual prior over states, but we define it as flat). The second is a version of the rational speech acts model of Frank & Goodman (2012) and Goodman & Stuhlmüller (2013), which has been shown to capture a broad range of SIs, but is known to be limited in its ability to derive manner implicatures and certain classes of embedded implicature (Bergen et al. 2012, 2014). The final two models are full versions of the one we presented in section 4. They represent opposite ends of the spectrum of non-trivial refinements. We saw in connection with table 3 and table 4 that there might be empirical value in greatly constraining the space of refinements.

We employ three methods of comparison: Pearson's correlation coefficient, which measures the linear correlation between the human responses and the model predictions; Spearman's rank correlation coefficient, which assesses how closely the human responses and model responses are aligned in terms of the rankings they predict; and the mean-squared error (MSE) of the model predictions as compared with the human responses, which summarizes the distance of the predictions from the human behavior. The use of these three measures allows us to assess which models

	Pearson		Spearman		MSE	
Literal semantics	.938	(.926 – .947)	.762	(.754 – .770)	.0065	(.0057 – .0075)
Fixed-lexicon pragmatics	.924	(.911 – .932)	.757	(.749 – .766)	.0079	(.0072 – .0090)
Unconstrained uncertainty	.945	(.936 – .950)	.794	(.767 – .820)	.0038	(.0035 – .0044)
Neo-Gricean uncertainty	.959	(.950 – .962)	.809	(.808 – .820)	.0034	(.0031 – .0040)

Table 5: Overall assessment with 95% confidence intervals obtained via non-parametric bootstrap over subjects.

best reproduce quantitative correspondence modulo arbitrary linear transformation (Pearson correlation), qualitative correspondence (Spearman correlation), and absolute fit between models and data. We find that the Spearman measure is often the most illuminating, since our fundamental goal is to reproduce the preference orderings revealed by the human responses. However, the three measures together yield a succinct multidimensional summary of how the models fare, and the same measures can be applied to particular target sentences to achieve more fine-grained insights.

Our model predictions are conditional probability distributions over states given messages, and hence constrained to be in the range $[0, 1]$ and to sum to 1. In contrast, our human responses are binary true/false judgments. To align these values, we rescale the human responses: if x^s is the 10-dimensional vector of percentage-true human responses for target sentence s, then each p^s is the vector of normalized values for that sentence, defined so that $p^s_i = x^s_i / \sum_{j=1}^{10} x^s_j$. This simply normalizes the responses into a conditional probability distribution over states given messages. The one noteworthy thing about this calculation is that, because it is done on a per-sentence basis, it is not a simple linear rescaling, and so it affects all of our assessment metrics when applied to multiple sentences at once. However, we regard it as the simplest viable linking hypothesis relating our model with our experimental data.

Figure 5 summarizes the models' predictions alongside the human responses. The predicted values are largely aligned for the examples without *some* in the object position. Where *some* occurs embedded, the models diverge in qualitative terms. For '*every*. . . *some*', the patterns are broadly similar, but only 'Neo-Gricean uncertainty' is able to mirror the preference ordering of responses seen in the human data. For '*exactly one*. . . *some*', only the two uncertainty models are able to predict local enrichment, in that only they assign high probability to the crucial worlds that are false on the literal construal: `NSA` and `SAA`. The 'Literal semantics' and 'Fixed-lexicon pragmatics' models are unable to predict the salience of these construals. Similarly, only the two uncertainty models predict '*none*. . . *some*' to have embedded enrichments leading to acceptability for `NNA`, `NAA`, and `AAA`. In broad strokes, we can say that 'Fixed-lexicon pragmatics' predicts only 'global' implicatures, those that CFS would obtain with unembedded exhaustification, whereas the two uncertainty models simulate embedded exhaustification (though without predicting it to be the most preferred option, in line with our human responses).

Table 5 summarizes our overall quantitative assessment. All of the correlations are extremely high, and the MSE values are extremely low. This is reassuring about the general utility of all of these models for predicting human judgments. In addition, the confidence intervals on the estimates are tight. We computed confidence in these estimates by a subject-wise non-parametric bootstrapping procedure, recomputing correlations for the same set of conditions, but with different simulated samples of participants. The resulting intervals reflect our confidence about estimates of

Figure 5: Analysis by target sentence, comparing model predictions with human responses.

these statistics for this particular set of experimental conditions.

Because of the high absolute values of all correlations, model comparison is important for interpretation. Two patterns stand out. First, 'Fixed-lexicon pragmatics' performs the least well overall. Since it has been shown to add substantial value in other areas of language and cognition, we conclude that its particular approach to enrichment is at odds with the patterns for embedded implicatures. The precise causes are hard to pinpoint, but the fact that our target implicatures are not always enrichments of the literal content is surely part of the problem. Second, neo-Gricean uncertainty achieves the best results across all three of our measures. Here again, this is consistent with our expectations based on the large illustrative example from section 4.4, where we saw that this constrained, lexically-driven approach to choosing refinements resulted in the best quantitative and qualitative pattern.

The overall analysis given in table 5 understates the value of both uncertainty models when it comes to the distribution of embedded implicatures. Our target sentences provide relatively little space for pragmatic enrichment; in figure 4, the left and middle columns essentially have only literal interpretations, leaving just the right column for our pragmatic models to shine. What's more, our qualitative review of figure 5 suggests that the right column examples reveal major distinctions. It's thus worth assessing them quantitatively in isolation. The results of such an assessment are in table 6. The most dramatic pattern is that the two fixed-lexicon models are unable to capture the patterns for embedded implicatures in the non-monotone and downward monotone environments. In contrast, both uncertainty models capture the patterns. These tight fits are evident in figure 5, and it is reassuring to see them reflected in our assessment measures.

It is striking that the literal model is competitive for '*every…some*'. This model does not distinguish among contexts in which the target sentence is true. Our participants only minimally distinguished among such readings, which makes sense in the context of a binary judgment task if we assume that the literal reading is accessible. However, the distinctions that do emerge from our experimental results align best with the preference-order predicted by the 'Neo-Gricean uncertainty' model, as revealed by the high Spearman coefficient.

Finally, it seems that neither uncertainty model is clearly superior to the other for these data: they are the top two models on all metrics, and are separated from each other by only a small amount. This suggests to us that we may have not yet found precisely the right approach to refinement. It is tempting to try additional refinement sets to find a single model that wins decisively for all the target examples. We are wary of doing this because, as noted above, it runs the risk of overfitting to our experimental responses; we could easily engineer our own success. However, this is nonetheless a fruitful avenue for future exploration if paired with additional experiments for further validation. Appendix A offers additional relevant findings.

Our model's performance is sensitive to the space of competitor messages, so it is worth asking how robust these findings are to changes in this area. We have found that the basic pattern is robust to a number of changes to the space of quantifiers. The only noteworthy finding we have to report in this regard is that allowing *only some* into object position has a major impact: while SSS remains the best-guess inference for the message '*every…some*' in this setting, '*exactly one…some*' and '*no…some*' effectively lose their embedded implicature readings. This makes intuitive sense given the nature of the model: if the speaker has the option to choose *only some of his shots*, and that form is equally costly, then surely her avoidance of that form in favor of *some of his shots* is a signal that she regards the local enrichment as infelicitous. As *only some* is made more costly, it becomes a less salient option, and embedded implicatures begin to reemerge.

	'every…some'			*'exactly one…some'*			*'no…some'*		
	P	S	MSE	P	S	MSE	P	S	MSE
Literal	.99	.86	.0002	.80	.70	.0180	.88	.52	.0346
Fixed-lexicon	.93	.85	.0027	.80	.70	.0179	.88	.52	.0346
Unconstrained	.88	.84	.0043	.98	.94	.0007	.76	.57	.0097
Neo-Gricean	.82	.88	.0087	.94	.87	.0036	.93	.89	.0028

Table 6: Assessment of crucial items. 'P' = 'Pearson'; 'S' = 'Spearman'.

7 Conclusion

With this paper, we sought a synthesis between Gricean accounts of pragmatic reasoning and grammar-driven ones like that of Chierchia et al. (2012). It seems to us inevitable that both grammar and interaction will play leading roles in the final theory of these phenomena; at some level, all participants in the debate acknowledge this. Our achievement is to unify the crucial components of these approaches in a single formal model that makes quantitative predictions.

The key components of the model we develop are compositional lexical uncertainty and recursive modeling of speaker and listener agents (Bergen et al. 2014). The lexical uncertainty property is in evidence in Chierchia et al.'s account as well, in the form of underspecified logical forms with context-dependent meanings. Our model has similar formal mechanisms but also offers an account of how discourse participants reason under this persistent linguistic uncertainty. This leads to an important conceptual point: not all underspecification has to be resolved in order for robust pragmatic enrichment to take place.

The recursive reasoning of our model is characteristic of both Gricean approaches and signaling systems approaches; our model shares formal properties of both but makes quantitative predictions of the sort that can be correlated with human preferences in communication. There are by now many models in the same family as ours (see, e.g., Camerer et al. 2004; Jäger 2012; Smith et al. 2013; Kao et al. 2014b; Jäger & Franke 2014), so further exploration is likely to yield an even more nuanced picture.

In addition, we saw that the space of refinements has a significant impact on the final predictions. It would thus be worthwhile to further explore different notions of refinement, seeking better fits with our own experimental patterns and then validating those conclusions in follow-up experiments using our experimental items, or applying the resulting models in new domains. For example, whereas refinement in the present model applies only to lexical entries, it could apply to phrases as well. Such phrasal refinements might be required to account for what Sauerland (2012, 2014) has called 'intermediate implicatures', where scalar strengthening seems to apply in between two (potentially non-monotonic) operators. However, study of the empirical distribution of such implicatures and the precise formal assumptions required to account for that distribution has only just begun. We have made publicly available all the data and code associated with this paper in an effort to encourage these and other new strands of theory development and quantitative assessment.

			$C(\mathbf{0})$	λ	k
Literal semantics	Pearson	.94			
	Spearman	.76			
	MSE	.0065			
Fixed lexicon pragmatics	Pearson	.93	1	.1	1
	Spearman	.76	0	.2	1
	MSE	.0069	1	.1	1
Unconstrained uncertainty	Pearson	.97	1	.1	1
	Spearman	.80	1	.1	1
	MSE	.0022	1	.1	1
Neo-Gricean uncertainty	Pearson	.98	1	.1	1
	Spearman	.81	1	.2	1
	MSE	.0018	1	.1	1

Table 7: Best models found in hyper-parameter exploration, as assessed against the binary-response experiment. The literal listener is not affected by any of the parameters explored.

A Parameter exploration

As we discussed in section 4.3, the definition of our model naturally suggests at least two extensions: (i) a temperature parameter λ modulating the speaker's inferences, and (ii) further iteration beyond the level of L. The full extended form of the model is defined as follows, again drawing on the objects and notational conventions established in section 4.3:

(20) a. $l_0(w \mid m, \mathcal{L}) \propto \mathcal{L}(m, w)P(w)$

b. $s_1(m \mid w, \mathcal{L}) \propto \exp\left(\lambda\left(\log l_0(w \mid m, \mathcal{L}) - C(m)\right)\right)$

c. $L_1(w \mid m) \propto P(w) \sum_{\mathcal{L}\in\mathbf{L}} P_{\mathbf{L}}(\mathcal{L}) s_1(m \mid w, \mathcal{L})$

d. $S_k(m \mid w) \propto \exp\left(\lambda\left(\log L_{k-1}(w \mid m) - C(m)\right)\right)$ (for $k > 1$)

e. $L_k(w \mid m) \propto S_k(m \mid w)P(w)$ (for $k > 1$)

From the perspective of this model, our decision to set $\lambda = 1$ and focus on L_1 might appear arbitrary. In addition, even from the perspective of our simpler model, our decision to fix the cost of the null message at 5 for all simulations and assessments was arbitrary. It is therefore worth exploring other settings for these hyper-parameters. To do this, we conducted a comprehensive grid search of the following values:

(21) a. λ: $[0.1, 2]$ in increments of .1, and $[3, 5]$ in increments of 1

b. L_k for $k \in \{1, 2, 3, 4, 5, 6\}$

c. $C(\mathbf{0}) \in \{0, 1, 2, 3, 4, 5, 6\}$

The grid search explores the full cross product of these values for each of our four models. For each setting, we conduct our standard model assessment against the data from our main (binary-response) experiment. Table 7 reports the best values for each of our four models, along with

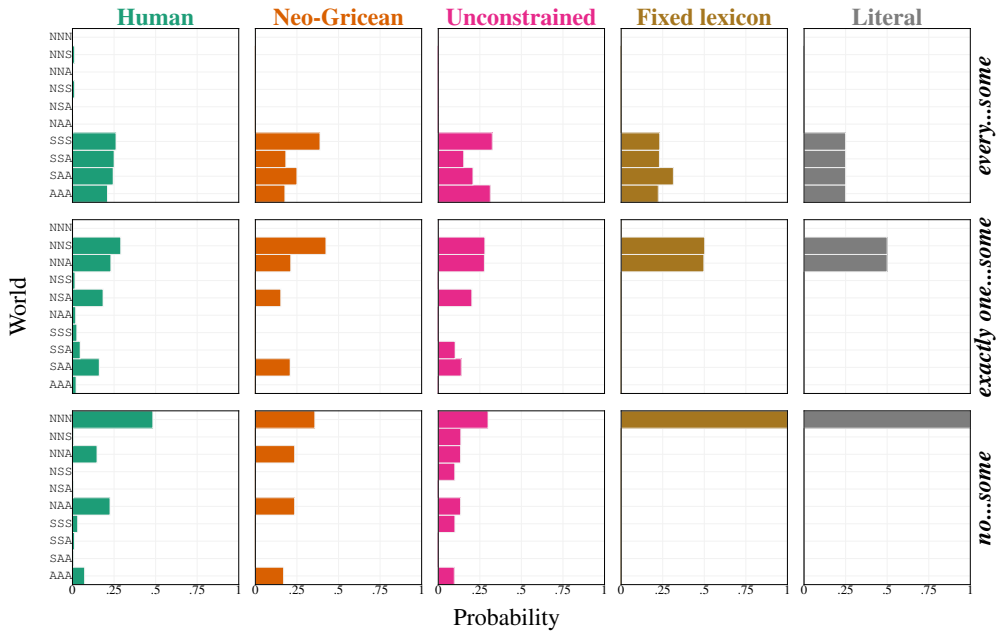

Figure 6: The crucial target sentences comparing the human data with L_1, using parameters in the range that seem to be nearly optimal for all of these models: $\lambda = 0.1$ and $C(\mathbf{0}) = 1$.

the minimal parameter settings that deliver those values. These results are consistent with our fundamental assessment of these models (section 6.3). Varying the cost of the null message has a relatively small impact on the outcomes, but the findings for the other two parameters may be relevant to broader discussions of *bounded* rationality in pragmatics. First, further iteration beyond L_1 is not necessary (Vogel et al. 2014). Second, the assumption in the main text that $\lambda = 1$, made primarily for clarity in deriving model predictions, does not provide the optimal fit to the experimental data: the value $\lambda = 0.1$ is slightly better. At lower values of λ, our listeners assume that speakers are paying little attention to the informativity of their messages, seeking only to be truthful (e.g., McMahan & Stone 2015). This is consistent with previous accounts according to which speakers are often unable to achieve ideal pragmatic calculations due to the the cognitive demands of production (Pechmann 1989; Levelt 1993; Engelhardt et al. 2006; Dale & Reiter 1995; van Deemter et al. 2012; Gatt et al. 2013). At the same time, the improvement is slight — compare table 7 to table 5 in the main text — and previous work has generally found that higher values of λ provide better predictions (for example, Kao et al. 2014a,b; Lassiter & Goodman 2015).

Figure 6 offers a finer-grained look at how these preferred settings affect outcomes for the crucial target items involving embedded *some*. The literal column is identical to the one in figure 5. The others are subtly different in ways that achieve a better match with the human data. For instance, the optimal parameters assign more probability to `AAA` in the '*no...some*' condition, which better matches the human responses. Overall, though, the contrasts between items are slightly dampened relative to the version of the model with $\lambda = 1$.

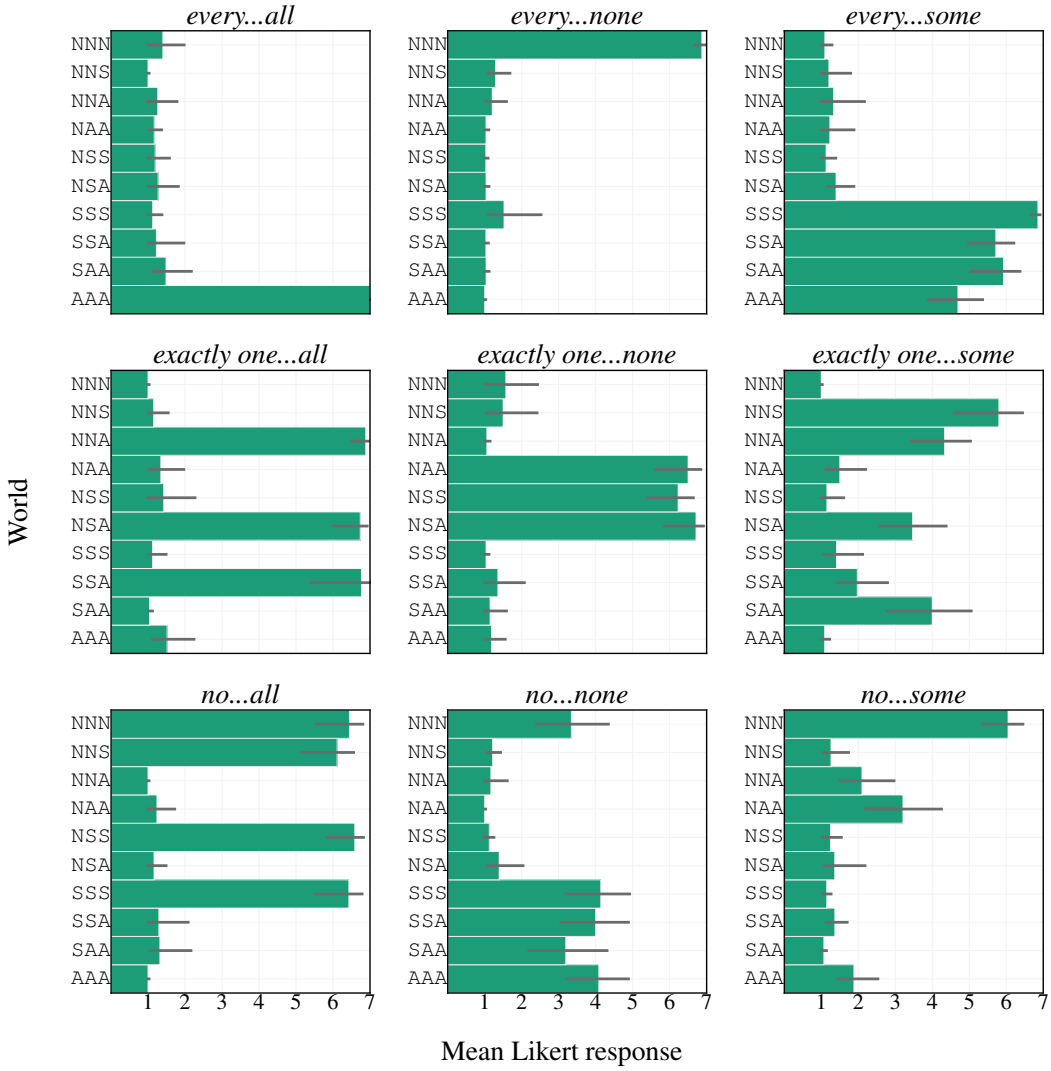

Figure 7: Likert-scale experimental results. Mean ratings by sentence with bootstrapped 95% confidence intervals.

B Likert-scale experiment

We conducted a version of the binary-response experiment discussed in section 6 using a Likert-scale for the response categories. Our rationale for using this scale was that it allows enough space for participants to both register a truth-value assessment and convey information about the quality of the report. This appendix reports briefly on this experiment. It yielded results identical in all important respects to those from our main experiment.

B.1 Methods

B.1.1 Participants

The experiment had 300 participants, all recruited with Amazon's Mechanical Turk. No participants or responses were excluded.

B.1.2 Materials

The displays were identical to those in figure 3, generated by the same procedures, but with the binary response categories replaced with a seven-point Likert scale ranging from 'Bad description' to 'Good description'. The target sentences were the ones in (15), and the conditions were as in (16). The same 23 fillers were used.

B.1.3 Procedure

After reading our consent form, participants were given the cover story in (17) with "judgments about the comments" replaced by "judgments about the quality of the comments". They completed the same three training items as were used in our main experiment. The design was again between-subjects. Each sentence received a total of 300 responses. For the target sentences, each sentence–world pair received between 19 and 44 responses (mean 30); this variation derives from our randomized procedure for assigning worlds to sentences.

B.2 Results

Figure 7 summarizes the responses by target sentence and world of evaluation. The results mirror those seen in figure 4 in all important respects. For our key theoretical comparisons, we again report significance levels using the nonparametric Mann–Whitney U test. In the '*every…some*' case, the highest ratings came in the SSS world. Worlds SSA and SAA received the next highest ratings (lower than SSS; both at $p < 0.001$). Of all the literally true worlds, AAA received the lowest rating (lower than SSA and SAA; both at $p < 0.05$). For the '*exactly one…some*' item, the highest ratings are again in the NNS condition, where it is true under its literal and locally enriched construals, but it also received high ratings in the two worlds where it is true only with local enrichment: NSA and SAA, which were both higher at $p < 0.05$ than in SSA, the world yielding the highest rating among those in which the sentence is false both literally and under all possible enrichments. As before, the strictly truth-conditional interpretation seems to be salient as well. Finally, we also find evidence for local enrichment under '*no…some*'. Condition NNN received the highest average ratings, suggesting a preference for a literal construal, but the ratings are high for the conditions requiring local enrichment: NNA, NAA, and AAA. The confidence intervals are wide, but a pooled comparison of {NNS, NSA} with {NNA, NAA, AAA} shows the latter set to be significantly higher-rated; $p = 0.006$.

B.3 Model assessment

Table 8 summarizes our model assessment. This assessment was done with identical settings and procedures to those reported in section 6.3, with one exception: since the minimal Likert value is 1,

	Pearson		Spearman		MSE	
Literal semantics	.935	(.910 – .947)	.756	(.742 – .764)	.0079	(.0065 – .0099)
Fixed-lexicon pragmatics	.920	(.894 – .932)	.751	(.736 – .759)	.0094	(.0080 – .0114)
Unconstrained uncertainty	.929	(.905 – .938)	.794	(.765 – .815)	.0052	(.0045 – .0067)
Neo-Gricean uncertainty	.950	(.927 – .956)	.805	(.795 – .812)	.0046	(.0038 – .0062)

Table 8: Overall assessment of the Likert-scale experiment with 95% confidence intervals obtained via by-subjects bootstrapping.

we subtract 1 from all scores when transforming them into the by-message normalized probability space of the model. Neo-Gricean uncertainty again emerges as the best model.

References

Alonso-Ovalle, Luis. 2008. Innocent exclusion in an alternative semantics. *Natural Language Semantics* 16(2). 115–128.

Bach, Kent. 1994. Conversational impliciture. *Mind and Language* 9(2). 124–162.

Bach, Kent. 2006. The top 10 misconceptions about implicature. In Betty Birner & Gregory Ward (eds.), *Drawing the boundaries of meaning: Neo-Gricean studies in pragmatics and semantics in honor of Laurence R. Horn*, 21–30. Amsterdam: John Benjamins.

Baker, C. L. 1970. Double negatives. *Linguistic Inquiry* 1(2). 169–186.

Beaver, David I. & Brady Zack Clark. 2008. *Sense and sensitivity: How focus determines meaning*. Oxford: Wiley-Blackwell.

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

Bergen, Leon, Roger Levy & Noah D. Goodman. 2014. Pragmatic reasoning through semantic inference. Ms., MIT, UCSD, and Stanford.

Blutner, Reinhard. 1998. Lexical pragmatics. *Journal of Semantics* 15(2). 115–162.

Bonawitz, Elizabeth, Stephanie Denison, Alison Gopnik & Thomas L Griffiths. 2014. Win-stay, lose-sample: A simple sequential algorithm for approximating Bayesian inference. *Cognitive Psychology* 74. 35–65.

Büring, Daniel & Katharina Hartmann. 2001. The syntax and semantics of focus-sensitive particles in German. *Natural Language and Linguistic Theory* 19(2). 229–281.

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

Chemla, Emmanuel. 2013. Apparent Hurford constraint obviations are based on scalar implicatures: An argument based on frequency counts. Ms. CNRS, ENS, LSCP Paris.

Chemla, Emmanuel & Benjamin Spector. 2011. Experimental evidence for embedded scalar implicatures. *Journal of Semantics* 28(3). 359–400.

Chierchia, Gennaro. 2004. Scalar implicatures, polarity phenomena, and the syntax/pragmatics

interface. In Adriana Belletti (ed.), *Structures and beyond: The cartography of syntactic structures*, vol. 3, 39–103. New York: Oxford University Press.

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

Chierchia, Gennaro, Danny Fox & Benjamin Spector. 2012. The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. In Maienborn et al. (2012) 2297–2332.

Clark, Eve V. & Herbert H. Clark. 1979. When nouns surface as verbs. *Language* 767–811.

Clark, Herbert H. 1997. Dogmas of understanding. *Discourse Processes* 23(3). 567–59.

Clifton, Charles Jr. & Chad Dube. 2010. Embedded implicatures observed: A comment on Geurts and Pouscoulous (2009). *Semantics and Pragmatics* 3(7). 1–13.

Dale, Robert & Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science* 19(2). 233–263.

van Deemter, Kees, Albert Gatt, Roger P.G. van Gompel & Emiel Krahmer. 2012. Toward a computational psycholinguistics of reference production. *Topics in Cognitive Science* 4(2). 166–183.

Degen, Judith. 2015. Investigating the distribution of *some* (but not *all*) implicatures using corpora and web-based methods. *Semantics and Pragmatics* 8(11). 1–55.

Degen, Judith & Michael K. Tanenhaus. 2015. Processing scalar implicature: A constraint-based approach. *Cognitive Science* 39(4). 667–710.

Engelhardt, Paul E., Karl G.D. Bailey & Fernanda Ferreira. 2006. Do speakers and listeners observe the Gricean maxim of quantity? *Journal of Memory and Language* 54(4). 554–573.

Fox, Danny. 2007. Free choice disjunction and the theory of scalar implicatures. In Sauerland & Stateva (2007) 71–120.

Fox, Danny. 2009. Too many alternatives: Density, symmetry, and other predicaments. In Tova Friedman & Edward Gibson (eds.), *Proceedings of Semantics and Linguistic Theory 17*, 89–111. Ithaca, NY: Cornell University.

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

Franke, Michael. 2009. *Signal to act: Game theory in pragmatics* ILLC Dissertation Series. Institute for Logic, Language and Computation, University of Amsterdam.

Gajewski, Jon. 2012. Innocent exclusion is not contradiction free. Ms., UConn.

Gatt, Albert, Roger P.G. van Gompel, Kees van Deemter & Emiel Krahmer. 2013. Are we Bayesian referring expression generators? In *Proceedings of the workshop on production of referring expressions: Bridging the gap between cognitive and computational approaches to reference*, Berlin.

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

Gazdar, Gerald. 1979b. A solution to the projection problem. In Choon-Kyu Oh & David A. Dinneen (eds.), *Syntax and semantics*, vol. 11: Presupposition, 57–89. New York: Academic Press.

Geurts, Bart. 2009. Scalar implicatures and local pragmatics. *Mind and Language* 24(1). 51–79.

Geurts, Bart. 2011. *Quantity implicatures*. Cambridge: Cambridge University Press.

Geurts, Bart & Nausicaa Pouscoulous. 2009. Embedded implicatures?!? *Semantics and Pragmatics* 2(4). 1–34.

Geurts, Bart & Bob van Tiel. 2013. Embedded scalars. *Semantics and Pragmatics* 6(9). 1–37.
Giles, Howard, Nikolas Coupland & Justine Coupland. 1991. Accommodation theory: Communication, context, and consequence. In Howard Giles, Nikolas Coupland & Justine Coupland (eds.), *Contexts of accommodation*, 1–68. Cambridge: Cambridge University Press.
Glucksberg, Sam. 2001. *Understanding figurative language: From metaphors to idioms*. Oxford University Press.
Goodman, Noah D. & Daniel Lassiter. 2015. Probabilistic semantics and pragmatics: Uncertainty in language and thought. In Shalom Lappin & Chris Fox (eds.), *The handbook of contemporary semantic theory*, Oxford: Wiley-Blackwell 2nd edn.
Goodman, Noah D. & Andreas Stuhlmüller. 2013. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science* 5(1). 173–184.
Grandy, Richard E. & Richard Warner. 2014. Paul grice. In Edward N. Zalta (ed.), *The stanford encyclopedia of philosophy*, Spring 2014 edn.
Grice, H. Paul. 1968. Utterer's meaning, sentence meaning, and word-meaning. *Foundations of Language* 4(3). 225–242.
Grice, H. Paul. 1975. Logic and conversation. In Peter Cole & Jerry Morgan (eds.), *Syntax and semantics*, vol. 3: Speech Acts, 43–58. New York: Academic Press.
Grice, H. Paul. 1989. *Studies in the way of words*. Cambridge, MA: Harvard University Press.
Griffiths, Thomas L, Falk Lieder, Noah D Goodman & Tom Griffiths. To appear. Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science* .
Grodner, Daniel J., Natalie M. Klein, Kathleen M. Carbary & Michael K. Tanenhaus. 2010. "Some," and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition* 116(1). 42–55.
Groenendijk, Jeroen & Martin Stokhof. 1984. *Studies in the semantics of questions and the pragmatics of answers*. Amsterdam: University of Amsterdam dissertation.
Hendriks, Petra, John Hoeks, Helen de Hoop, Irene Krammer, Erik-Jan Smits, Jennifer Spenader & Henriette de Swart. 2009. A large-scale investigation of scalar implicature. In Uli Sauerland & Kazuko Yatsushiro (eds.), *Semantics and pragmatics: From experiment to theory*, 30–50. Houndmills, Basingstoke, Hampshire: Palgrave Macmillan.
Hirschberg, Julia. 1985. *A theory of scalar implicature*. Philadelphia: University of Pennsylvania dissertation.
Horn, Laurence R. 1972. *On the semantic properties of logical operators in English*. Los Angeles: UCLA dissertation.
Horn, Laurence R. 1984. Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In Deborah Schiffrin (ed.), *Meaning, form, and use in context: Linguistic applications*, 11–42. Washington, D.C.: Georgetown University Press.
Hurford, James R. 1974. Exclusive or inclusive disjunction. *Foundations of Language* 11(3). 409–411.
Ippolito, Michela. 2010. Embedded implicatures? Remarks on the debate between globalist and localist theories. *Semantics and Pragmatics* 3(5). 1–15.
Israel, Michael. 1996. Polarity sensitivity as lexical semantics. *Linguistics and Philosophy* 19(6). 619–666.
de Jager, Tikitu & Robert van Rooij. 2007. Explaining quantity implicatures. In *Proceedings of the 11th conference on theoretical aspects of rationality and knowledge*, 193–202. New York:

ACM Digital Library.
Jäger, Gerhard. 2007. Game dynamics connects semantics and pragmatics. In Ahti-Veikko Pietarinen (ed.), *Game theory and linguistic meaning*, 89–102. Amsterdam: Elsevier.
Jäger, Gerhard. 2012. Game theory in semantics and pragmatics. In Maienborn et al. (2012) 2487–2425.
Jäger, Gerhard & Michael Franke. 2014. Pragmatic back-and-forth reasoning. In Salvatore Pistoia Reda (ed.), *Pragmatics, semantics and the case of scalar implicature*, 170–200. Houndmills, Basingstoke, Hampshire: Palgrave Macmillan.
Kao, Justine T., Leon Bergen & Noah D. Goodman. 2014a. Formalizing the pragmatics of metaphor understanding. In *Proceedings of the 36th annual meeting of the Cognitive Science Society*, 719–724. Wheat Ridge, CO: Cognitive Science Society.
Kao, Justine T., Jean Y. Wu, Leon Bergen & Noah D. Goodman. 2014b. Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences* 111(33). 12002–12007.
Lascarides, Alex & Ann Copestake. 1998. Pragmatics and word meaning. *Journal of Linguistics* 34(2). 387–414.
Lassiter, Daniel & Noah D. Goodman. 2013. Context, scale structure, and statistics in the interpretation of positive-form adjectives. In Todd Snider (ed.), *Proceedings of semantics and linguistic theory 23*, 587–610. Ithaca, NY: CLC Publications.
Lassiter, Daniel & Noah D. Goodman. 2015. Adjectival vagueness in a Bayesian model of interpretation. *Synthese* .
Levelt, Willem J.M. 1993. *Speaking: From intention to articulation*, vol. 1. MIT Press.
Levinson, Stephen C. 2000. *Presumptive meanings: The theory of generalized conversational implicature*. Cambridge, MA: MIT Press.
Lewis, David. 1969. *Convention*. Cambridge, MA: Harvard University Press.
Lewis, David. 1970. General semantics. *Synthese* 22(1). 18–67.
Magri, Giorgio. 2009. A theory of individual-level predicates based on blind mandatory scalar implicatures. *Natural Language Semantics* 17(3). 245–297.
Maienborn, Claudia, Klaus von Heusinger & Paul Portner (eds.). 2012. *Semantics: An international handbook of natural language meaning*, vol. 3. Berlin: Mouton de Gruyter.
Marr, David. 1982. *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: WH Freeman and Company.
McCawley, James D. 1978. Conversational implicature and the lexicon. In Peter Cole (ed.), *Syntax and semantics*, vol. 7: Pragmatics, 245–259. New York: Academic Press.
McMahan, Brian & Matthew Stone. 2015. A Bayesian model of grounded color semantics. *Transactions of the Association for Computational Linguistics* 3. 103–115.
Muskens, Reinhard. 1995. *Meaning and partiality*. Stanford, CA: CSLI/FoLLI.
Paris, Scott G. 1973. Comprehension of language connectives and propositional logical relationships. *Journal of Experimental Child Psychology* 16(2). 278–291.
Pechmann, Thomas. 1989. Incremental speech production and referential overspecification. *Linguistics* 27(1). 89–110.
Potts, Christopher & Roger Levy. 2015. Negotiating lexical uncertainty and speaker expertise with disjunction. In *Proceedings of the 41st annual meeting of the Berkeley Linguistics Society*, Berkeley, CA: BLS.
Reed, Ann M. 1991. On interpreting partitives. In Donna Jo Napoli & Judy Anne Kegl (eds.), *Bridges between psychology and linguistics: A Swarthmore festschrift for Lila Gleitman*, 207–

223. Hillsdale, NJ: Erlbaum.
Rooth, Mats. 1985. *Association with focus*. Amherst, MA: UMass Amherst dissertation.
Rooth, Mats. 1992. A theory of focus interpretation. *Natural Language Semantics* 1(1). 75–116.
Rooth, Mats. 1996. Focus. In Shalom Lappin (ed.), *Handbook of contemporary semantic theory*, 271–298. London: Blackwell.
Russell, Benjamin. 2006. Against grammatical computation of scalar implicatures. *Journal of Semantics* 23(4). 361–382.
Russell, Benjamin. 2012. *Probabilistic reasoning and the computation of scalar implicatures*. Providence, RI: Brown University dissertation.
Sanborn, Adam N., Thomas L. Griffiths & Daniel J. Navarro. 2010. Rational approximations to rational models: alternative algorithms for category learning. *Psychological Review* 117(4). 1144–1167.
Sauerland, Uli. 2001. On the computation of conversational implicatures. In Rachel Hastings, Brendan Jackson & Zsofia Zvolenszky (eds.), *Proceedings of Semantics and Linguistic Theory 11*, 388–403. Ithaca, NY: Cornell Linguistics Circle.
Sauerland, Uli. 2004. Scalar implicatures in complex sentences. *Linguistics and Philosophy* 27(3). 367–391.
Sauerland, Uli. 2010. Embedded implicatures and experimental constraints: A reply to Geurts & Pouscoulous and Chemla. *Semantics and Pragmatics* 3(2). 1–13.
Sauerland, Uli. 2012. The computation of scalar implicatures: Pragmatic, lexical or grammatical? *Language and Linguistics Compass.* 6(1). 36–49.
Sauerland, Uli. 2014. Intermediate scalar implicatures. In Salvatore Pistoia Reda (ed.), *Pragmatics, semantics and the case of scalar implicatures*, 72–98. Basingstoke: Palgrave MacMillan.
Sauerland, Uli & Penka Stateva (eds.). 2007. *Presupposition and implicature in compositional semantics*. Houndmills, Basingstoke, Hampshire: Palgrave Macmillan.
Schulz, Katrin & Robert van Rooij. 2006. Pragmatic meaning and non-monotonic reasoning: The case of exhaustive interpretation. *Linguistics and Philosophy* 29(2). 205–250.
Smith, Nathaniel J., Noah D. Goodman & Michael C. Frank. 2013. Learning and using language via recursive pragmatic reasoning about other agents. In *Advances in neural information processing systems 26*, 3039–3047.
Spector, Benjamin. 2007a. Aspects of the pragmatics of plural morphology. In Sauerland & Stateva (2007) 243–281.
Spector, Benjamin. 2007b. Scalar implicatures: Exhaustivity and Gricean reasoning. In Maria Aloni, Paul Dekker & Alastair Butler (eds.), *Questions in dynamic semantics*, 225–249. Amsterdam: Elsevier.
Sperber, Dan & Deirdre Wilson. 1995. *Relevance: Communication and cognition*. Oxford: Blackwell 2nd edn.
Stiller, Alex, Noah D. Goodman & Michael C. Frank. 2011. Ad-hoc scalar implicature in adults and children. In Laura Carlson, Christoph Hoelscher & Thomas F. Shipley (eds.), *Proceedings of the 33rd annual meeting of the Cognitive Science Society*, 2134–2139. Austin, TX: Cognitive Science Society.
Sutton, Richard S. & Andrew G. Barto. 1998. *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.
van Tiel, Bob. 2014. Embedded scalars and typicality. *Journal of Semantics* 31(2). 147–177.
Vogel, Adam, Andrés Gómez Emilsson, Michael C. Frank, Dan Jurafsky & Christopher Potts.

2014. Learning to reason pragmatically with cognitive limitations. In *Proceedings of the 36th annual meeting of the Cognitive Science Society*, 3055–3060. Wheat Ridge, CO: Cognitive Science Society.

Vul, Edward, Noah Goodman, Thomas L Griffiths & Joshua B Tenenbaum. 2014. One and done? Optimal decisions from very few samples. *Cognitive Science* 38(4). 599–637.

Wilson, Dierdre & Robyn Carston. 2007. A unitary approach to lexical pragmatics: Relevance, inference and ad hoc concepts. In Noel Burton-Roberts (ed.), *Pragmatics*, 230–259. Basingstoke and New York: Palgrave Macmillan.