A Minimal Parsimonious Model of Pragmatic Comprehension

Larry Muhlstein

July 21, 2016

University of California San Diego

9500 Gilman Drive

La Jolla, CA 92092

Abstract

Computational accounts of how we infer what a speaker means by what she says have recently spurred significant progress in our understanding of linguistic pragmatics. These models, motivated by general assumptions about human cognition, have been shown to account for a wide range of phenomena in pragmatic comprehension. Despite their successes, these models incorporate a number of structures and assumptions that diminish their parsimony as well as their consistency with the complex cognitive features of human communication. In this paper, I demonstrate these weaknesses through a number of empirical test cases and theoretical arguments. I then propose an alternative computational modeling approach, derived from the popular Rational Speech Act (RSA) theory, that makes fewer and simpler assumptions about pragmatic comprehension and accounts for a wide range of pragmatic phenomena without requiring ad hoc modifications or theoretically unmotivated assumptions. This Speaker Centric Inference Model of Pragmatic Interpretation (SCIMPI) not only provides the basis for a simple and parsimonious account of human pragmatic inference, but suggests that formal linguistic pragmatics should be practiced by incrementally constructing theory through the introduction of minimal and motivated components.

Linguistic Pragmatics

In his seminal paper, *Logic and Conversation*, H.P. Grice had the insight that a speaker will choose the utterance that is likely to be most informative to the listener about what they wish to communicate and that the listener can use this knowledge in order to infer what the speaker meant by what she said (Grice, 1975). This insight engendered modern linguistic pragmatics, the study of how agents use context to make inferences and resolve ambiguity when communicating. All contemporary theories of pragmatics include some mechanism by which a listener makes inferences about the intentions of the speaker. Typically, such theories lean on a set of normative maxims that specify how a speaker should act when communicating. By combining these maxims (typically either Grice's original maxims or Horn's (Horn, 1984) condensed taxonomy) with knowledge of a speaker and what was said, an agent can infer what that speaker was likely to have meant.

Most of these theories also employ mechanisms of counterfactual inference whereby a listener draws conclusions about what the speaker meant by comparing what was said with what the speaker might have said but chose not to. If the speaker follows the assumed normative maxims, then the speaker's choice of a particular utterance is informative about her not intending to mean something that would have caused her to choose a different utterance. This of course requires that listener have knowledge of the set of things that the speaker might intend to communicate (henceforth referred to as the alternative meaning set) as well as knowledge of the set of utterances that the speaker chose her utterance from (henceforth referred to as the alternative utterance set). However, as humans, we do not have access to the minds of others and we therefore have at best limited knowledge of these sets. Pragmatic inference also requires that we have some information

about how people use language to communicate. This knowledge about the relationship between meanings and utterances is also limited. Despite these epistemic constraints, contemporary accounts of pragmatic phenomena assume that the listener has full knowledge of these alternative sets and the mappings between them and proceed to derive the phenomena of interest from there[1]. Because of the pervasiveness of this uncertainty, I argue that a compelling theory of pragmatics must be able to represent the graded nature of these beliefs, and should be able to derive empirically observed pragmatic phenomena from computations performed on these uncertain beliefs. In order to more fully motivate the need for a new theory, we first explore the space of formal theories of pragmatics and then delve into the details of the current leading edge models to see where they succeed and where theoretical modifications might help us to better understand how people determine what is meant by what is said.

Formal Theories of Pragmatics

Formal theories of pragmatics capture the structure of the theories in the form of mathematical formulae in order to allow for precise representation and prediction of empirical observations. This predictive precision enables the theories to be effectively supported or rejected using experimental data without the possibility for post hoc reinterpretation of theoretical predictions given new data. Formal theories take the shape of models, which are in effect reifications of theoretical hypotheses and assumptions in terms of mathematical structures. There are two major subclasses of formal pragmatic models, those which represent language as a logical system, and those whose representations take the form of computationally defined functions.

Logical theories

Most of the literature on formal theories of pragmatics has focused on developing what I will call logical theories of pragmatics. These theories usually take the form of either predicate or propositional logic with the addition of specialized operators such as those representing exhaustification or modality. Logical theories, however, are incapable of representing graded uncertainty. While an epistemic or other modal logic may have the ability to represent the existence of uncertainty, it cannot capture relative degrees of this uncertainty. In order to do this, we need a probability calculus. In addition, logical theories have difficulty specifying predictions in terms of data that might actually be observed. They operate on variables and predicates and though these may refer to observable features of language, neither are directly reflective of human judgements that might be elicited by experiment. Because of this feature, logical theories of language are not much more falsifiable than informal (ex: maxim-based) theories, as their validation or falsification depends not only on the precise structure of the theory but on how evidence is informally interpreted before it meets the theory itself. Because of both of these features, some contemporary researchers of linguistic pragmatics have began to formulate their theories in terms of computational models.

Computational theories

Instead of using logical formulae, computational models of pragmatics structure their predictions in terms of mathematical functions that represent the computations that they claim are being performed by humans using language. The word "computational" here is used in two specific senses: that the model defines a computation that is being performed by the agent and that the model is defined at Marr's computational level of abstraction. The latter usage is important for computational models of pragmatics in that it allows them to specify predictions about human behavior whilst being relatively agnostic about the processes by which humans perform these computations. This idea comes from the work of David Marr (1982), who argued that models of cognition can be structured at three different levels of abstraction, the implementation level, the algorithmic level, and the computational level. Roughly, the implementational level refers to the actual hardware structures that are performing the computations, the algorithmic level specifies the process by which these computations are being performed, and the computational level specifies only the functional relationship between the input data, the parameters, and the output data that represent the computation

[1]While there do exist a few theories of pragmatics that do not make this assumption, they are either insufficiently precise or insufficiently general to solve the open questions at hand. I will discuss a few of these theories as they becomes relevant in the text.

being performed. Specifying a model of pragmatic inference at the computational level allows the model to predict what humans will do without having to worry about how they do it. Contemporary models of pragmatic inference are all defined at the computational level.

The two classes of contemporary computational models of pragmatics are the iterated best response (IBR) (Jäger, 2008; Franke, 2013) and the rational speech act (RSA) models (Frank and Goodman, 2012). The IBR model is formalized in the language of game theory whereas the RSA model is in the language of Bayesian probability. Both of these model classes capture roughly the same insights: that pragmatic inference can be represented in terms of computations on a set of utterances and a set of meanings (usually referents), and that recursive reasoning about interlocutor choices derives pragmatic inferences about the meaning that the speaker probably intended to convey by choosing a particular utterance.

Because they capture similar behavior, I will discuss only one class of models throughout this paper. I will use the RSA framework since RSA models are probabilistic and the issues that I wish to discuss involve graded knowledge and uncertainty.

The recursive model structure of RSA is intended to capture a rational type of inference based on counterfactuals–the other choices that the interlocutor might have made. Based on Gricean pragmatic theory, or Horn's (1984) simplification of it, the model assumes that the speaker will always choose an utterance so as to be maximally informative to the listener. With this assumption, and with a model of utterances as having literal meanings defined by mappings from the set of possible utterances to the set of possible meanings, the speaker is able to choose the utterance most likely to convey a particular meaning given her model of the listener's inference process.

One of the key insights of the RSA and IBR models of pragmatics is that the speaker's model of the listener can take many possible forms. In the base or "literal listener" case, the speaker's model of the listener consists only of these "semantic" mappings from the set of utterances to the set of meanings. If this is the process by which the speaker makes decisions about which utterances to produce, then the optimal listener will model this speaker and then choose an interpretation based on which meaning the speaker would have been most likely to intend to convey given the utterance they produced. As we will see, this mechanism is able to derive scalar implicatures and other forms of pragmatic inference from the circumstances of the counterfactual choices of the speaker. However, if the speaker knows that the listener is using this kind of model for interpretation, she can incorporate this information into her production decisions. In this way, the speaker might make use of a more sophisticated model of the listener in order to be more informative. However, the buck does not stop here. The listener in turn could chose to adopt this model of the speaker, and the speaker could chose to adopt this even more complex model of the listener, and so on ad infinitum. This recursive simulation of your interlocutor's choices (assuming that he is performing a recursive simulation of you as well), is what allows these recursive counterfactual models of pragmatics to derive their predictions of human pragmatic inference.

RSA Formal Model Specification

Formally, RSA is structured in terms of set-theoretic language model and probability distributions that interact with that language model. The language–or "lexicon" as they call it—is specified as a set of discrete mappings between a set of possible meanings and a set of possible utterances. The meaning set M is usually defined by the context of possible referents presented to the participants, and is therefore restricted by the experimental design. The set of possible utterances U is less well-defined and is usually generated by the researchers to reflect the set of utterances that they believe the speaker is likely to produce in the current scenario. The mappings between meanings and utterances that comprise the lexicon are discrete and binary. This means that each utterance has a set of meanings that it may be used to convey and a set of meanings that it cannot be used to convey. Because the lexicon is defined only in terms of these binary mappings, the model does not represent a graded lexical semantics. A word or utterance[2] can only mean or not mean, refer or not refer. Additionally, the lexical specification used in RSA is non-directional, which means that it does not specify an injection or a surjection from one set to another, but an arbitrary mapping. We will see

[2]In this paper I use "word" and "utterance" relatively interchangeably. All of the models considered here do not process compositional linguistic structure, but simply operate on discrete linguistic forms. The example forms used in most experiments and simulations are often single words, but they need not be. The models therefore make predictions about "utterances," but the formal representations of these utterances are unstructured.

some consequences of this decision later in the paper. Lexica are implemented as a matrix where the rows correspond to utterances and the columns correspond to meanings. Each entry in this matrix is either a 1 or a 0 depending on whether it is admissible that the utterance and the meaning map to each other. An example lexicon matrix can be seen in table 1.

			
"Blue"	1	1	0
"Green"	0	0	1
"Square"	1	0	1
"Circle"	0	1	0

Table 1: Example RSA lexicon

In addition to this lexicon, RSA uses a set of equations that specify how this lexical knowledge is used to derive pragmatic inferences. These equations are recursively defined and begin with a literal listener as the base case as seen in equation 1.

$$l_0 = ln(lex(u, m)) \tag{1}$$

This literal listener l_0 is defined in terms of the lexicon $lex(u, m)$–a mapping between utterances $u \in U$and meanings $m \in M$– and the natural logarithm function, which gives the informativeness of each utterance about each meaning using the information theoretic concept of negative surprisal. The first order speaker S_1 uses this listener model in order to decide which utterance to produce. The speaker takes the distribution over utterances and meanings given by the literal listener and performs a Bayesian inversion (treating the literal listener as a conditional distribution over meanings given utterances) to yield a conditional distribution over utterances that the speaker might choose to produce given the meanings that the speaker might want to convey. The speaker then takes this distribution and applies a parameterized softmax function, which allows the speaker's decision behavior to move between the optimal case of always choosing the utterance most likely to convey the intended meaning to the literal listener and the agnostic case of choosing each utterance with equal probability. This decision strength parameter α is often fit to the data, but in the initial RSA paper (Frank and Goodman, 2012) it is set to $\alpha = 1$ in order to simulate probability matching behavior or a Luce choice decision rule. The speaker model can be seen in equation 2 where $P_{S_0}(u|m)$ refers to the first order speaker probability distribution over utterances given meanings.

$$P_{S_1}(u|m) \propto \exp(\alpha l_0) = lex(u, m)^\alpha \tag{2}$$

The first order listener uses this first order speaker model in order to decide on which meaning the speaker probably intended to convey given the utterance that the speaker chose to produce. This listener model also takes into account the prior probabilities $P(m)$ of the speaker actually wanting to communicate each meaning. This listener model appears as in equation 3.

$$P_{L_1}(m|u) \propto P_{S_1}(u|m)P(m) \tag{3}$$

In the instantiations of core RSA that we explore in this paper, this is the extent of the model. However, because the first order speaker and listener are defined in terms of lower order speakers and listeners, this model can be expanded recursively to include a second order speaker and listener, a third order speaker and listener, etc. without adding any other modeling structure. This full recursive specification of RSA can be seen in equations 4-5 where the literal listener l_0 appears as specified above.

$$P_{S_n}(u|m) \propto P_{L_{n-1}}(m|u)^\alpha \tag{4}$$

$$P_{L_n}(m|u) \propto P_{S_n}(u|m)P(m) \tag{5}$$

Using this recursive formulation, RSA can represent counterfactual pragmatic inference at arbitrary depth (a listener reasoning about his model of the speaker reasoning about her model of the listener reasoning about his model of the speaker, etc.), as long as there is full mutual knowledge between the speaker and the listener.

How RSA Accounts For Some Empirical Phenomena

Ad-hoc Scalar Implicature

The RSA model of pragmatics is primarily motivated by the case of scalar implicature and handles it cleanly. A clear example of how RSA derives scalar implicatures comes from the Frank and Goodman's initial paper defining the model (Frank and Goodman, 2012).

In their example, a human participant is given a set of colored shapes (see figure) that a speaker might be talking about. They are given the prompt, "Imagine someone is talking to you and uses the word 'blue' to refer to one of these objects. Which object are they talking about?"

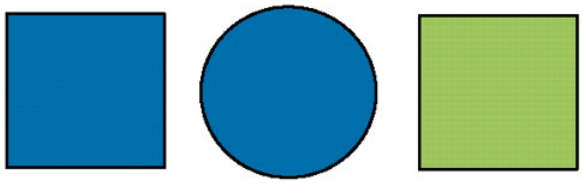

Figure 1: An example set of potential referents presented to participants in the initial RSA study.

The participants overall favored the blue square, but some also chose the blue circle. The RSA model derives these predictions by modeling the fact that both the blue circle and the blue square are consistent with the word "blue," but that the speaker, had she wanted to refer to the blue circle, could have uniquely referred to it using the term "circle," or had she wanted to refer to the green square, could have uniquely referred to it using the term "green." Therefore if the speaker decided to say "blue," and if we assume that the speaker is trying to be as informative as possible, a rational listener would infer that the speaker used the most informative term that she could and therefore wanted to refer to the blue square. The prior probability of the speaker choosing each referent is also taken into account. Frank and Goodman measured this empirically and found that the participants believe the blue square to be likely half as likely to be referred to a priori as each of the other two referents. The relationship between their model predictions and their empirical data can be seen in figure 2.

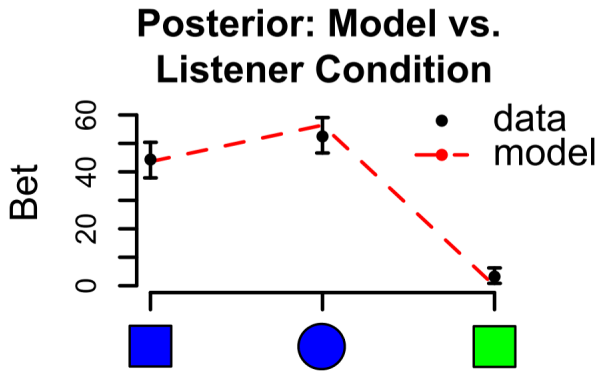

Figure 2: RSA model predictions and empirical data for ad hoc scalar implicatures (from Frank and Goodman (2012)).

This basic insight that pragmatic inference can be modeled by capturing counterfactual decisions made on the basis of informativity has been successfully extended to a variety of domains. Extensions of the core RSA model have been able to capture non-literal use of number words (Kao et al., 2014b), metaphorical language use (Kao et al., 2014a), revision of world knowledge (Degen et al., 2015), specificity implicatures and horn implicatures (Bergen et al., 2012), uncertainty about your speaker's lexicon (Bergen et al., 2014), and even word learning (Frank and Goodman, 2014). The breadth of this applicability suggests that this model

class is capturing something important about human pragmatic reasoning and should be taken seriously. However, the application of RSA to each of these domains makes use of additional computational structures beyond the core RSA model. It may simply be that these are inherently more complex phenomena. In order to account for complex things, one might expect to have to add additional structure to the model. While such structure is often necessary and important, too much unmotivated structure can affect the ability of a model to generalize. In the following two scenarios, we examine what needs to be added to this core RSA model in order to account for the class of specificity implicatures and the effects of relevance present in a more realistically sized language.

Horn/Specificity Implicature

Another important class of pragmatic inferences are what are known either as specificity implicatures or Horn implicatures. These were first noticed by Horn (1984), and are believed to derive from the simple idea that the more specific the utterance is, the more informative it would be to a listener, and therefore a speaker trying to be maximally informative would choose the most specific utterance available. This means that a speaker who chooses an utterance that is less informative than an alternative utterance for a given meaning M is not likely to have meant to convey M at all. Instead, she is more likely to have meant to convey an alternative meaning for which this utterance would have been maximally informative. The most famous case of specificity or Horn implicature[3] comes from the contrast between the utterances "some" and "all." In the case where a speaker might be taken to be choosing her utterances between "some" and "all" (and possibly others such as "none" that also fall on this lexical scale), "some" often comes to have a pragmatically enhanced meaning. "Some" is taken by most semanticists to have a "literal semantics" of the form $\exists x, cat(x) \Rightarrow cute(x)$ (informally corresponding to the claim that at least one of any cats that might exist is cute) (for the example sentence "Some cats are cute.") However, the sentence "Some cats are cute" is often read as having the meaning $\exists x : cat(x) \Rightarrow cute(x) \wedge \exists y : cat(y) \wedge \neg cute(y)$ (informally corresponding to the claim that some but not all cats are cute) instead. This interpretation is often taken as a pragmatic strengthening of the literal semantics based on a specificity implicature[4]. The claim is that the listener infers that if the speaker knew that all cats were cute, then she would have used the word "all" instead of "some" because "all" is more specific and therefore more informative. Given that she instead said "some cats are cute," she must not have been in a position to use the stronger alternative "all" and so she must have meant that some but not all cats are cute. Since this story is much more complex if we take into account the full landscape of logical and grammatical theories of how this phenomena operates, we narrow our focus to the RSA framework and how it proposes to account for general specificity implicatures.

In order to account for this type of pragmatic inference, Bergen, Levy, and Goodman added additional structure to the core RSA model, yielding what they call the lexical uncertainty model (Bergen et al., 2014). In this model, they consider the possibility that the speaker might be producing language with uncertainty about which lexicon the listener is actually using for interpretation. To do this, they introduce the lexicon as an auxiliary random variable with multiple possible values preset for the scenario. They then run the core RSA model over each lexicon and merge the resulting predictions by summing or marginalizing over the values obtained from each lexicon. In some variants, additional RSA iterations are run yet again on these merged predictions. Formally, this can be viewed as in equation 6.

$$P(m|u) = \sum_{lex \in LEX} RSA(lex, u)p(lex) \tag{6}$$

Where $RSA(lex, u)$ refers to the computations defined by the vanilla RSA model as described in equations 1-5, LEX refers to the set of all lexica considered by the first order speaker, and $p(lex)$ denotes the prior probability of a specific lexicon. In most uses of this model in the literature, LEX does not contain all possible lexica, but some subset of them, and $p(lex)$ is usually chosen as the uniform distribution over $lex \in LEX$.

In order to derive the "some" vs. "all" specificity implicature, Bergen et al. (2014) use the set of lexica seen in figure 3.

[3]These terms are mostly used interchangeably in the literature with a few exceptions. They are used interchangeably in this paper except where explicitly noted otherwise.

[4]Although a few authors, most notably Chierchia et al. (2008) claim that this is the literal semantics of "some."

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

Figure 3: Lexica used in the Bergen et al. (2014) lexical uncertainty account of the "some" vs. "all" specificity implicature.

How this gives rise to the specificity implicature shown in figure 4 is best understood by looking at the model predictions for each lexicon individually, and then considering how they sum together. Applying RSA to $\mathcal{L}_1$ yields a basic scalar implicature where "all" refers only to $\forall$ and where "some" is more likely to refer to $\exists\neg\forall$ but may also refer to $\forall$. Applying RSA to $\mathcal{L}_2$ also yields a categorical result where "all" refers only to $\forall$ and where "some" refers only to $\exists\neg\forall$. Applying RSA to $\mathcal{L}_3$ yields another categorical distribution where "all" and "some" both refer only to $\forall$. Note that $\mathcal{L}_3$ could only be computed by adding a "null utterance" or u_{null} to the lexicon to avoid division by zero, since it contains no utterances that can possibly refer to $\exists\neg\forall$. Summing over the outcomes for each lexicon yields the distribution seen in figure 4 where "all" can refer only to $\forall$ and where "some" has some probability of referring to $\forall$, but is most likely to refer to $\exists\neg\forall$, which is the empirically observed result. Note also that the only lexicon altered by running RSA was $\mathcal{L}_1$. Running RSA on $\mathcal{L}_2$ and $\mathcal{L}_3$ yielded only the original mappings as specified in the lexicon. Additionally, the implicature is generated only when this set of possible lexica is chosen explicitly by the researchers. If they were to perform this inference on the set of all possible lexica (without some non-uniform prior distribution over the lexica), the model would predict that each utterance has an equal probability of referring to each referent, giving rise to the "symmetry problem[5]."

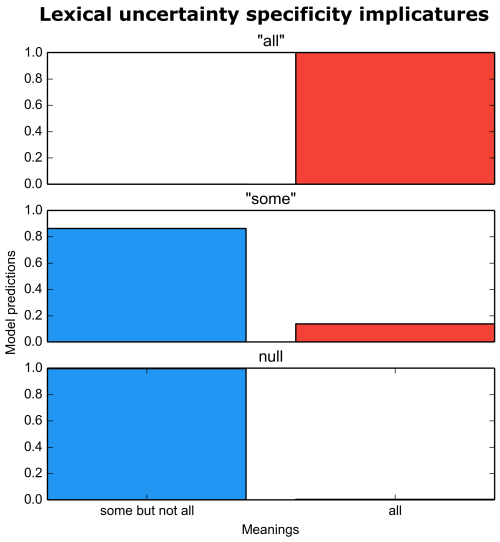

Figure 4: Predictions of the lexical uncertainty model for the "some" vs. "all" case of specificity implicature.

Relevance effects and expanding the language

The final scenario against which we will evaluate the RSA model concerns itself with how sensitive the model is to an expansion of linguistic knowledge to a more realistic representation. Though this is not really discussed in the formal literature because it is taken for granted that pragmatic theories should do this, this "sensitivity to alternatives" comprises an important set of tests for any theory of pragmatics. All of the

[5]See Bergen et al. (2014) for details.

literature on RSA has focused on testing the model in well-defined scenarios with very small meaning and alternative utterance sets. While this has allowed for the experimental verification of the model performance in these strictly delimited test cases, the question remains as to whether RSA is able to make similarly effective predictions when given much larger and therefore more realistic meaning and alternative utterance sets.

The key concern is that the addition of auxiliary meanings and referents will dilute the predictions of the model such that they no longer reflect the crisp pragmatic judgements made by human comprehenders in situ despite their large vocabularies and sets of available meanings. An initial look at this problem asks two questions: how do the model predictions change when the set of alternative utterances is expanded? and how do the model predictions change when the set of alternative meanings is expanded?

Both of these concerns are addressed informally by a branch of the pragmatics literature called "relevance theory (Sperber and Wilson, 1986)." Relevance theory accounts for a number of pragmatic and communicative phenomena based on the ideas that people tend to communicate only about "relevant" content, and that people have a (shared) mechanism by which they determine what utterances and meanings are relevant to the current scenario. If speakers and listeners both agree upon the sets of relevant alternative meanings and utterances, then they need not consider all others, and they can perform efficient computations that give rise to strong pragmatic predictions. If a listener assumes that the speaker is producing only relevant content, then the resultant presuppositions allow for strong pragmatic inferences about what the speaker actually means by what she says. Notably, all contemporary work on relevance theory takes place in the informal domain. This is because it is unclear how to determine which utterances and meanings are relevant without considering all of the utterances and meanings in the first place. This problem has led to recent work on "context coordination," performed by Mollica et al. (2015) and by Muhlstein et al. (2015).

In order to handle these concerns, we would like a pragmatics model to be able to increase the size of its vocabulary and possible meanings such that it is able to perform strong implicatures over many different possible meanings and utterances due to the differential contextual relevance of the meanings and utterances. In other terms, we would like to be able to represent additional possible but irrelevant meanings and utterances without affecting the predictions made in strictly delimited cases. To test this, we will attempt to increase the size of the alternative utterance sets and the alternative meaning sets independently and see if it is possible to maintain strong ad hoc implicatures.

Before delving headfirst into simulation, we might note that the prior distribution over meanings in the RSA model is capable of modulating the relevance of each meaning such that meanings with lower prior probabilities do not affect the resultant predictions. By altering this distribution, additional meanings may be added with low prior probabilities that will not significantly affect the model predictions. RSA has therefore passed the first test (sensitivity to the size of the meaning set) without us having to run a single model.

The second test, however, may be more difficult. RSA does not include a prior distribution over utterances, nor could it be simply appended with one (due to renormalization of the utterance probabilities by the speaker and the listener's observation of the actual utterance produced). Additionally, the specification of the language using a set-theoretic language model does not allow for flexible representation of the probabilities that utterances may be used to communicate meanings. Figure 5 shows the effects of adding additional alternative utterances, "Regular Quadrilateral," and "Bluesquare" to the RSA model, and how they affect the predictions seen in figure 2.

Here we can see that the predictions of the basic RSA model are notably affected by the addition of only two alternative utterances. Because there is no way to represent the prior probability of each utterance, there is no way for the model to capture the relevance of each utterance to the current scenario or speaker. It could be that the current speaker is very unlikely to say either "Regular Quadrilateral" or "Bluesquare" and that these utterances should therefore not play a noticeable role in the model predictions, but RSA has no way of representing this property. A model of linguistic pragmatics should not require that the comprehender only have knowledge of relevant and common words, but should also allow the comprehender to consider and interpret less relevant and rare words. We expect that such words should have little effect on the model if not said, but should drive a significant implicature if the speaker happens to say them despite their perceived irrelevance. Neither of these properties can be represented by vanilla RSA.

The lexical uncertainty model, however, does not have this problem. By considering many possible lexica, each of which the speaker may be modeling a literal listener as using for interpretation, it is possible

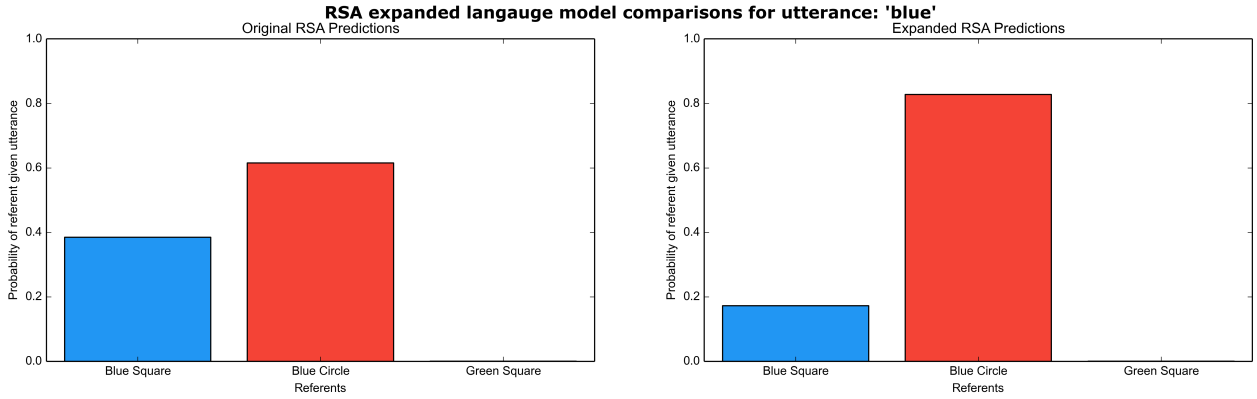

Figure 5: Comparison of RSA ad hoc predictions as seen earlier with those that arise from the representation of an expanded language. The utterances "Regular Quadrilateral" and "Bluesquare" were added, with the first having the same extension as the word "square" and the second referring only to the blue square referent.

to capture the effects of differential likelihood or relevance of utterances. Though this seems to save the RSA family in this scenario, there are problems with this approach too, but they will be detailed in the next section.

An Alternative Approach: The Speaker-Centric Interpretation Model of Pragmatic Inference (SCIMPI)

Motivation

Building off of the insights offered by RSA, we can derive an alternative model that accounts for all of the phenomena discussed above without having to add any additional structure ad hoc. RSA made use of the idea that pragmatic inference could be captured by a rational model of human decision making (Frank and Goodman, 2012), but it is not the only way in which this powerful approach might account for linguistic pragmatic inference. In this section, I articulate an alternative model of pragmatic inference, the Speaker-Centric Interpretation Model of Pragmatic Inference (SCIMPI[6] for short) model[7], which builds upon the power of the RSA approach in order to generalize further and allow for a simpler and more natural computational-level theory of pragmatic inference.

The extent of uncertainty

RSA is unique in the set of contemporary theories of pragmatics in that it represents graded uncertainty. It models the listener's uncertainty about the speaker's knowledge state, the speaker's uncertainty about the listener's knowledge state, and so on recursively. Variants of RSA represent additional uncertainty in their auxiliary structures. The lexical uncertainty model represents uncertainty about the lexicon that the speaker is making use of (strictly speaking, it's uncertainty in the listener's beliefs about the speaker's beliefs about the lexicon that the listener is using for interpretation.) Other RSA variants represent uncertainty about speaker affect and literal vs. figurative language use Kao et al. (2014b, 2013, 2014a) and other unobserved

[6]There is an intended double entendre here, for the partial orthographic and strong phonological similarities between SCIMPI and "skimpy" suggest that this model does not include anything that is unmotivated or unnecessary to account for the core phenomenon of pragmatic inference. While the word "skimpy" certainly has a connotation that suggests an insufficient amount of stuff, I do not believe this to be the case and instead take this as a humorous suggestion. I welcome all critics of my approach to make use of this suggestive connotation in their introductions so as to provide a bit of humor at what would otherwise be my expense.

[7]I am aware of the redundant use in "SCIMPI model" and do not believe this to be grammatically problematic for reasons of non-compositional linguistic theories that I hold and would be willing to discuss over drinks with any interested or offended parties.

quantities, which they represent as "lifted variables"[8] Goodman and Lassiter (2014). These models do not, however, allow for the full extent of uncertainty present in a listener's interpretation judgements. While an argument might be made that this is due to a focus on specific phenomena and a desire to keep the model simple, there are certain kinds of uncertainties that are difficult to reasonably represent in RSA and which have a strong effect on pragmatic interpretation processes.

To motivate this claim, let us take a closer look at the representation of uncertainty about strict linguistic knowledge in the RSA family of models. By strict linguistic knowledge I simply mean knowledge the listener holds about the way in which language is used to convey meanings, not knowledge about the relationships between various contextual or "real-world" factors that may affect this use of language. Such factors can be treated as auxiliary (but still really important!) for the purposes of this analysis. In the core RSA model, there is no uncertainty about the nature of the language. Instead, the language is represented as a series of mappings between two sets, the set of meanings and the set of utterances. If there exists a mapping between an utterance and a meaning, then that utterance is capable of denoting that meaning. If there does not exist such a mapping, the utterance is not capable of denoting that meaning. Because of this representation of lexical knowledge as domain-codomain mappings, all utterances that are capable of denoting a given meaning have the exact same probability of doing so a priori. This means that any asymmetries that make one utterance more likely than another to convey a meaning must come from differences in the number of meanings that might be conveyed by an utterance (the prior informativity of the utterance). These factors do not, however, allow for a general representation of uncertainty in lexical knowledge. It is not possible, for instance, to represent the fact that a speaker is more likely to use the word "mug" to talk about the receptacle holding my coffee than the word "demitasse," even though the word demitasse is strictly more specific and is capable of referring to my coffee-carrying container. Because of this, is also impossible for two words with the same extensions (same set of allowed meanings) to have different posterior probabilities. For example, if the words "soda," and "soda-pop" have the same meanings associated with them, RSA is not capable of representing the fact that one word might be more a priori likely to be produced than the other.

In order to handle these problems, Bergen et al. (2014) introduced the notion of lexical uncertainty, which allows for the representation of uncertainty about which lexicon the speaker is using. Such a model could handle the "demitasse" example by explicitly representing multiple lexica that the speaker believes the listener might be considering. One lexicon might include the word demitasse and another lexicon might not, and altering the prior probability that the speaker is considering each interpretation lexicon would modulate the posterior probability that the speaker is considering the word "demitasse." It could handle the "soda-pop" example by representing two lexica where one contains the word "soda" and not the word "soda-pop," and where the other is the opposite. Modulating the prior probabilities of these lexica would allow the model to represent different likelihoods of these words (however, it is important to note that these words would not be competing against each other as scalar alternatives and therefore would not be pragmatically relevant to each other.)

Such a modification does not come without issues of its own, however. For example, the nature of the specific lexica that the listener's model of the speaker is considering affects the types of uncertainties that can be represented. If there are only two lexica where one represents a more learned listener and the other a more colloquial listener, and if the learned listener model contains not only the additional word "demitasse," but also other words such as "amanuensis" and "murenger," then the model cannot represent different amounts of uncertainty about the use of "demitasse" without also affecting the probability that speaker's model of the listener contains knowledge of obscure words for dictation and wall-making professions. This potentially problematic non-independence of lexical uncertainties can be avoided by representing all possible lexica containing all possible utterances and meanings and mappings between them, but the size of such a knowledge store would expand super-exponentially in both the number of meanings and utterances and is therefore highly cognitively implausible. In fact, the number of entries that need to be represented in such a model is a *super-exponential* function of *both* the number of meanings and the meaning set and the number of utterances in the alternative utterance set. The computations required by such a model would scale equally astronomically. Furthermore, the linguistic knowledge would be represented both in terms of this set of lexica and in terms of a probability distribution over these lexica given the context. Such a configuration is not reflective of any evidence we have about how humans represent their lexical knowledge.

[8]Technically, these lifted variables are simply additional random variables in the model about which no posterior inference is made. Their values are weighted by prior distributions that are input as parameters to the models.

This representational decision is designed to make the model work, which is acceptable, but strictly weaker than a representational decision with more empirical plausibility all other things being equal.

Instead of representing uncertainty in terms of a distribution over set mappings, the SCIMPI model represents a conditional probability distribution of utterances given meanings. Such a distribution allows for a listener to hold knowledge of how likely a speaker is to say U given that she wants to convey M in a manner that is independent of other things a speaker might mean or say. The size of this knowledge representation scheme grows only with the product of the number of meanings and the number of utterances and is capable of representing everything that a full lexical uncertainty model does. Moreover, the interdependencies between these pieces of knowledge are not directly part of the core pragmatics model, but may come from separate representational structures that derive these interdependencies from theories about other cognitive processes. If you had a theory that people who know the word "demitasse" are also more likely to know the words "amanuensis" and "murenger," then you could build such a theory on top of this core pragmatics model and it would allow you to represent these interdependencies, however SCIMPI does not implicitly contain such a theory in its model structure (as does lexical uncertainty), nor does it require any theory of structured lexical interdependencies at all. Additional structure of this sort naturally compliments the SCIMPI model; it does not constrain it. SCIMPI is therefore more cognitively plausible because it is able to derive pragmatic predictions effeciently, and without the need to posit latent structures whose cognitive existence is dubious at best.

Speaker-centricity

Another important choice in the design of SCIMPI involves a decision about how the form of the listener's linguistic knowledge is represented. The set-theoretic manner that the RSA model uses to represent language knowledge is not directly reflective either of the kind of information that a listener uses for interpretation, nor of the kind of knowledge that a speaker uses for production, but is rather a sort-of theoretical idealization of the notion of a language. In any pragmatics model that yields graded judgements, the listener's ultimate decisions can be expressed in terms of a distribution over meanings given utterances, and the speaker's decisions similarly make use of a distribution over utterances given meanings. Instead of working directly with linguistic knowledge of this form, RSA computes this knowledge from it's set-theoretic language representation and a set of structured recursive computations on this knowledge. Specifically, RSA represents the listener's beliefs about the speaker's beliefs about the listener's model of the language and uses this knowledge to compute the listener's predictions for what the speaker means by what she says. By choosing this representational scheme and this computational structure, RSA makes an implicit claim that this is the kind of language knowledge that a listener uses to make pragmatic inferences. While there is no prima facie difficulty with this representational choice, it does mean that RSA requires some additional theoretical machinery to account for how language might be learned and used. For example let's say that Kelly, a young child, encounters her mother saying "ball" and pointing to a particular object (one that we might call a "ball"). Assuming that Kelly is able to infer that her mother intends to call that object a "ball[9] ," the form of the knowledge Kelly receives is not simply the co-occurrence of the word "ball" and the ball object, but the fact that her mother, in wanting to communicate about the object, used the word "ball." This knowledge is therefore not in the form of a set mapping, but is a conditional distribution of utterances given meanings.

The SCIMPI model takes this production distribution to be the primitive form of linguistic knowledge. It claims that listeners compute implicatures and other pragmatic inferences by taking their beliefs about what their interlocutor is likely to say given what she means ($p(u|m)$), applying a prior distribution ($p(m)$) over what she might mean, and using Bayesian inversion to compute what the speaker is likely to mean given what she says ($p(m|u)$). These beliefs are, of course, sensitive to the context and to the particular interlocutor, but the distribution need not be stored rote. It can be computed using other world and contextual knowledge structures. The SCIMPI model simply claims that this speaker-centric knowledge, whether it comes from direct prior experience with the interlocutor, from experience with similar interlocutors, or from knowledge

[9]There is actually a lot of empirical evidence that suggests not only that Sally might be able to do this, but also that this plays a big role in how she learns the word "ball." Due to this evidence, we will not consider the myriad mereological problems that one might encounter here such as the philosohpically-motivated inability to distinguish between the ball and the lower right half of the ball or between the ball and the ball and the wooden block sitting next to it. Assume that all humans share some notion of "object" and have consistent object-detection mechanisms.

of what you would say in order to convey something, is the form of the linguistic knowledge that goes into pragmatic inference.

At face value, the treatment of linguistic knowledge as production-centric might be the most controversial claim made by the SCIMPI model specification. However, the model is robust to a few ways in which this assumption might not be strictly true. For example, this production centric knowledge may be straightforwardly computed from other kinds of linguistic knowledge including from a joint distribution over meanings and utterances (assuming that Kelly is only getting co-occurrence information), or from listener-centric knowledge (assuming that listeners store their knowledge directly based on what they have previously found to be meant given what was said). In either of these cases, the relationship between the predictions of the SCIMPI model and these variants is dependent solely on the manner in which the speaker-centric knowledge is said to be computed from the other knowledge form. This being said, there is evidence that strong sampling assumptions (Tenenbaum and Griffiths, 2002; Tenenbaum, 1999) and the Bayesian size principle allow children to gain implicit negative evidence from the kinds of observations they make when acquiring language simply because they take into account the fact that there is a whole wealth of possible data that they didn't observe(Hsu and Griffiths, 2009). This suggests that children do acquire language more in terms of conditional probabilities than joint distributions. Furthermore, a model built directly on listener-centric knowledge would have difficulty accounting simultaneously for the facts that people do compute implicatures and other pragmatically enhanced meanings and that speakers do not seem to model listeners to a strong degree(Lane and Ferreira, 2008). To capture the implicature property in a model with listener-centric knowledge, the listener would need to model a speaker modeling a listener, since direct use of listener-centric knowledge does not result in any counterfactual inferences. Any such model would suggest that the speaker is also using strongly pragmatically enhanced inferences for production and would therefore be incompatible with the second requirement that speakers should be largely greedy. Though this argument does not rule-out the possibility that we are primarily making use of such knowledge nor that we are using such knowledge to some smaller degree, it does suggest that a speaker-centric knowledge representation is a good place to start.

Recursive structure

Another motivation for the SCIMPI model comes from a set of observations about the recursive structure present in the RSA model. This structure derives from the economic theory of rational decision making and makes direct use of results from Thomas Shelling's work on focal points (Schelling, 1960) and John Maynard Keynes' work on recursive "beauty contests" (Keynes, 2006). These theories build on game theoretic assumptions and argue that a rational agent should make a decision based on the maximization of their own utility given what actions they expect their partner or competitor to take. In order to predict what actions their partner or competitor will take, this requires that the rational agent have a model of this other agent. However, it is not so simple. If we take into account that the other agent (let's call him agent B) in this scenario also has a model of his interlocutor (let's call him agent A), then agent B's model of agent A must take into account agent A's model of agent B. This creates recursive interdependence between the models. A's model of B and B's model of A cannot be simultaneously accurate, since each model depends on the other model. In such a scenario–assuming that all the relevant knowledge is shared and accurate–the best prediction is made by an agent who models their interlocutor perfectly, which means that this agent performs exactly one more recursive step than the other agent. Importantly, Shelling showed that, in a cooperative game, the limit of these recursions converges to the mutually best solution (or strategy) to the problem (the focal point.) RSA uses this result to suggest that optimal communicative behavior comes from the application of this recursion to the maximal depth that is feasible to the agents. This provides uncertainty as to the optimal depth of recursion, a problem with may be ameliorated by appeal to resource-limited accounts of rational behavior (Griffiths et al., 2015) where the expected performance benefit of additional recursion is directly traded for the cognitive costs of performing this additional computation.

This story, while formally justifiable in the case of perfect mutual knowledge, becomes much more complicated when the strength and accuracy of each interlocutor's beliefs about each other is taken into account. Real agents are not only limited in terms of their cognitive computational resources, but also in terms of their knowledge. In the simplest case of linguistic pragmatics, this knowledge takes the combined form of the prior probabilities of what the speaker will want to communicate ($p(m)$) and the probabilities of what

the speaker will say given what they want to communicate. However, this "knowledge" is actually just a set of beliefs that may include significant uncertainty and that may be incorrect. Incorporation of these beliefs into a model requires the unpacking of a substantial amount of structure that is hidden by the assumption of strong mutual knowledge. For example, to model a listener reasoning about a speaker reasoning about a listener with uncertainty, you need to separately represent the listener's beliefs about the speaker's beliefs about the listener's interpretation function, the listener's beliefs about the speaker's beliefs about the listener's beliefs about the prior probability distribution over what the speaker intends to communicate, and the listener's beliefs about the prior probability distribution over what the speaker intends to communicate. More realistically, you might not want to assume that the speaker is modeling a listener, but might also want uncertainty over whether the speaker is taking the listener into account, since it is unclear how much the speaker might be thinking about her utterances. If you want to represent this possibility as well, then you will need an additional probability distribution over the listener's beliefs about the speaker's production function and another distribution over the expected recursion depth that each interlocutor model is taking into account. And this is only for two steps of recursion[10]! The number of distributions that need to be represented to compute this full model of recursive inference under uncertainty grows exponentially with the depth of the recursion. Even worse, this knowledge needs to be acquired somehow. It seems highly unlikely that a listener has enough data to be able to acquire strong beliefs about his interlocutor's beliefs about his own interpretation distribution since this nested information is only observable in the most indirect of forms. The acquisition of knowledge about further recursively nested hypothetical interlocutors is neigh impossible. While these facts are true for a model wishing to represent all of this uncertainty, a model built off of some additional assumptions might not need to represent all of this knowledge. For example, one might have a theory that humans simplify this knowledge by cognitively conflating (or equating) their beliefs about their interlocutor's beliefs about them with their own beliefs about themselves. This would provide a biased but simplified system from which inferences could be made. However, there are many possible such theoretical simplifications and there is no way to adjudicate between them except via empirical data, which was not used in the RSA model. Essentially, under realistic assumptions about interlocutor uncertainty, recursive models of pragmatic inference such as RSA[11] are not nearly as well-motivated or elegant as they appear under the assumption of strong mutual knowledge.

Instead of representing full recursive uncertainty, the SCIMPI model represents a summary distribution of the listener's beliefs about what the speaker will say given what she wants to communicate. This distribution captures general pragmatic inference without requiring a specific theory about recursive reasoning. A recursive theory may be plausible if there is full mutual knowledge, but in the real-world case where such knowledge does not exist–and where there is no strong reason to believe that the speaker is acting by modeling some sort of diluted listener– the most prima facie plausible theory is one where the listener uses a simpler set of beliefs about what the speaker is likely to say given what she wants to convey. Such a model allows the listener to make use of his linguistic knowledge in the form in which it is naturally acquired and it greatly reduces the required number of theoretical assumptions about the nature of human pragmatic interpretation. Moreover, such a non-recursive or "flattened" model of comprehension allows for more specific theories to "plug in" and influence the production knowledge structure. For example, someone might have a theory that claims that the production knowledge is built from recursive interlocutor modeling in special cases of high mutual knowledge and where there is high importance placed on correct interpretation. In these cases, this theory can be used to generate the SCIMPI model's production knowledge distribution using another set of knowledge, while still retaining the simpler model for other scenarios. In this way, this non-recursive knowledge specification is strictly more general than the recursive RSA equivalent in that it allows for a broad range of domain-specific theoretical structures to interface in a modular manner. This allows SCIMPI to generalize further than RSA and to be less dependent on the validity of strong normative assumptions such as that of mutual knowledge.

[10]The amount of recursion that RSA models typically represent

[11]As well as variants of RSA like the lexical uncertainty model.

The Model

Formally, the SCIMPI model takes the form of a simple Bayesian inversion of the speaker-centric knowledge weighted by a prior distribution over meanings. This can be seen in equation 7.

$$p_L(m|u) \propto p_S(u|m)p_S(m) \tag{7}$$

Here $p_L(m|u)$ is a discrete probability distribution capturing the listener's inferences about what the speaker probably means m by the utterance $u$she produced. It takes the form of the posterior distribution in Bayes' rule and therefore represents the model's predictions. The likelihood $p_S(u|m)$ is another discrete probability distribution that represents the listener's beliefs about the speaker's production. It is given to the model either through direct empirical elicitation of these probabilities or by computing them from other models that capture additional theory about how the listener infers what a speaker is likely to say. In the simplest case, these probabilities may be gleaned through direct word-learning and represent a summary of the knowledge about language use that the listener has so far acquired. The prior distribution $p_S(m)$ captures the beliefs of the listener about how likely the speaker is to try to convey each meaning. It also might be elicited directly via experiment or computed from auxiliary models.

We might also want to be more explicit about the dependencies involved here, since the ways in which a speaker uses language vary and the prior probabilities of what she might want to communicate are not always the same. To represent this, we add a random variable c to stand in for context. Since all predictions are made given a context, and since the SCIMPI model does not include a theory about the effects of context on the likelihood and prior distributions, we simply add the context as a conditioning variable in each distribution. This can be seen in equation 8[12].

$$p_L(m|u,c) \propto p_S(u|m,c)p_S(m|c) \tag{8}$$

This set of equations gives rise to the same counterfactual inference dynamics as the RSA model because the alternative utterances that a speaker might have produced give rise to an explaining-away type effect where the speaker is not likely to have produced an utterance. This falls directly out of the Bayesian inversion because the alternatives essentially "compete" for probability mass. If the speaker is very likely to say X given that she wants to convey M, but she actually said Y, then the listener will be able to infer that the speaker is not likely to have meant to convey M because if she did she would have said X and not Y. In this way, the listener's knowledge of what the speaker actually said interacts with his knowledge of what she might say given her intentions in order to produce a distribution over what she probably did intend to communicate. It turns out that this simple dynamic, upon interaction with different structures of uncertainty, gives rise to a broad set of pragmatic phenomena. We will see examples of this in the next section.

There is nothing particularly complicated about the SCIMPI, model, and in a certain sense it is more of a theoretical contribution than a modeling one. The framing of this theory as a computational model, however, allows its predictions to be borne out in a precise and parametric manner. The strong claims about the relationship between each model component and an empirical phenomenon allow for the direct comparison between these model predictions and real world phenomena. In the following section, I review the phenomena introduced in the RSA section and demonstrate how SCIMPI fares in each domain.

Behavior on the previously identified empirical phenomena

In the section on RSA, we considered what kinds of structural modifications were necessary for the model to account for a few important classes of empirically observed pragmatic phenomena. In this section, we pit the SCIMPI model against the same phenomena and examine the necessary and sufficient conditions for it to capture the same empirical findings.

[12]Note that the SCIMPI model does not require the parameterized softmax transformation (as seen in RSA) in order to capture the speaker's decision strength. This is because the SCIMPI model represents the uncertainties explicitly and therefore captures the "greediness" of the speaker inside of the production distribution itself.

Ad hoc Scalar Implicature

Ad hoc scalar implicature, the raison d'être of the core RSA model, is the most central case of pragmatic inference and therefore the first test that a model of pragmatics must face. The SCIMPI model is able to capture scalar implicature in a similar manner to RSA, but without the need for recursion. Ad hoc implicatures are derived from the fact that the speaker might have said something other than what she actually said, and the assumption that the speaker chose what she actually said because it was more informative than the other possibilities. The SCIMPI model takes direct advantage of this counterfactual process and represents it explicitly as a Bayesian inference about what the speaker probably meant by what she said, given information about what the speaker is likely to say to convey each meaning and about how likely the speaker is to intend to communicate each meaning. Because this inference is Bayesian, it naturally takes advantage of information about other utterances that the speaker might have alternatively produced in an "explaining away" manner.

Experiment: Empirical evaluation of SCIMPI on the case of ad hoc implicature

Since the SCIMPI model requires a speaker-centric knowledge distribution and predicts that this knowledge–when run through SCIMPI–gives rise to the expected pragmatic inferences, it should be the case that 1. people actually have this knowledge in some form, and 2. people's speaker-centric knowledge and their prior beliefs about communicative intentions $P(m)$ predict their pragmatic judgements about a speaker's production.

Methods

In order to evaluate the SCIMPI model on the case of ad hoc implicature, I designed a simple experiment to elicit participants' speaker-centric knowledge as well as their beliefs about the prior and posterior probabilities of reference to each object in a given context. The experimental design was based largely on the Frank and Goodman (2012) study. I used the same context of possible referents (a blue square, a blue circle, and a green square) as well as similar prompts to elicit both prior and posterior probability judgements. To capture the participants' speaker-centric knowledge, I asked each participant to rate the likelihood that John, the imaginary interlocutor, would produce each word in order to refer to each of the objects. These prompts appeared to the participants as shown in figure 6.

The order of presentation of all stimuli and the order of presentation of likelihood judgement prompts were both randomized. All judgements were measured by providing the participants with a slider that ranged from 0-100 labeled both with these numbers at intervals of 10 as well as with the labels "impossible," "completely unsure," and "certain" at the beginning, midpoint, and end of the scale respectively. As there was no constraint that the participants' ratings add up to 100, the measurements were normalized to sum to 100 within participants before any further data processing was performed[13].

Results

60 participants were recruited via Amazon Mechanical Turk. They were each compensated in accordance with UCSD IRB #141149. One subject's results were excluded because they were not a native speaker of English, and the results of another subject were excluded because they noted that they made a mistake in their answers. This yielded a final n of 58.

Data was analyzed in a between-subjects manner by computing the means and 95% confidence intervals for the normalized participant judgements. These mean normalized priors and likelihoods were then used as the parameters of the SCIMPI model. Figure 7 shows the empirical priors and the empirically-measured likelihoods that correspond to the speaker-centric knowledge of the participants.

The empirically measured priors were qualitatively but not quantitatively consistent with those shown in the Frank and Goodman (2012) RSA study. For the analysis of the SCIMPI model, we used the priors elicited in our experiment, but for comparisons between models (as shown later in this paper), we used the same priors for both models, somewhat arbitrarily choosing to use the priors from the RSA study. Figure

[13]This was done both so that the judgements could be used as proper probabilities as well as to ensure that each participant's judgments were considered equally.

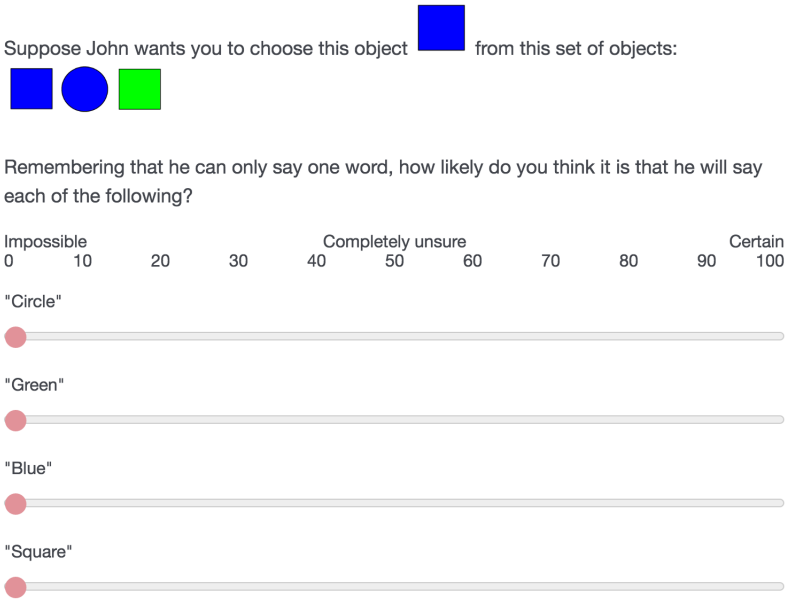

Figure 6: An example prompt asking participants to provide their speaker-centric knowledge about the blue square referent.

8 juxtaposes the predictions of the SCIMPI model with the posterior judgements made by participants in our task. Overall, there is a very good qualitative fit between model and data, with notable quantitative discrepancies only in the probability that the speaker's production of the utterance "blue" refers to the green square. This discrepancy is likely to be an artifact of the way judgements were measured, as the scales that were provided to participants do not lend themselves to accurate estimation of small probabilities.

This gives rise to the right hand plot in figure 9, which qualitatively matches the RSA model predictions for an underlying distribution reflective of a noisy version of the lexicon (i.e. where the word "blue" maps mostly to blue objects with a small probability of mapping to a green object and so on for the other words in the alternative utterance set.)

Taking a look at the two plots, it is worth noticing that the SCIMPI model places a small probability on the utterance "blue" referring to the green square. This is desirable because it reflects the fact that the listener is not entirely certain about what the speaker tends to say in order to communicate. In this case, the probability that "blue" refers to the green square might come to represent the probability that the speaker misunderstands the words in the language, that the listener misunderstands the words in the language, that the listener mishears the speaker, and so on. In cases where words are more ambiguous than shape and primary/secondary color words, this representation of communicative uncertainty is even more essential. Base RSA is unable to represent this uncertainty because it has a fixed lexicon. In order to represent uncertainty about the language using the RSA model, we need to invoke the lexical uncertainty alternative. Because SCIMPI generates ad hoc implicatures using fewer theoretical assumptions than RSA and because it captures more realistic properties of the language such as uncertainty, it may reasonably be said to provide a better account of the phenomenon.

Horn/Specificity Implicature

Next on the list of empirical phenomena we looked at with RSA is specificity implicature. Since there were no empirical data to work with, I had to determine how to best and most justifiably set the SCIMPI model parameters so as to test whether it plausibly generates specificity implicatures in a manner consistent with our expectations. I handled this in two ways. First, I wanted to see if a simple algorithmic approach to merging the candidate lexica from the RSA lexical uncertainty model could give rise to a set of mappings that, when run through the SCIMPI model, yield reasonable predictions. To do this, I took the lexica that Bergen

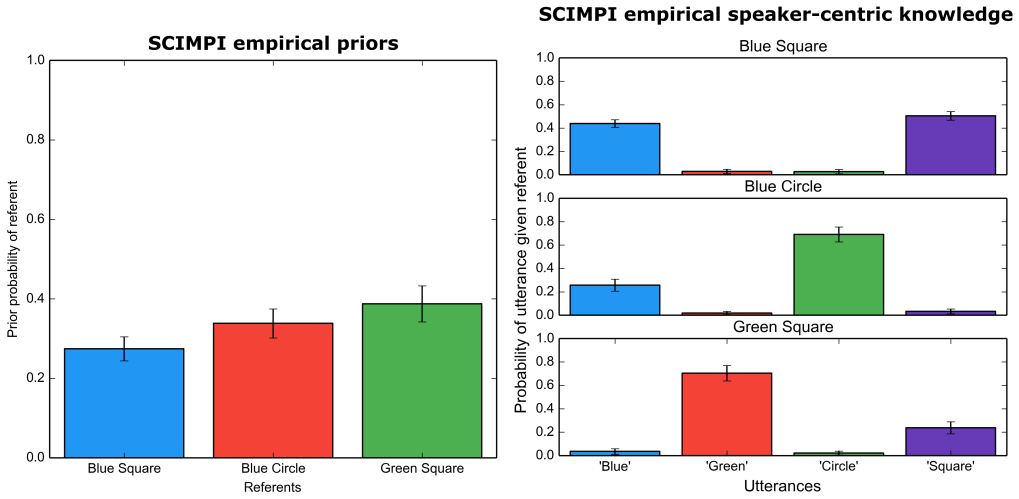

Figure 7: The left plot shows the empirically measured priors whose values correspond qualitatively, but not quantitatively, to the priors measured in the original RSA study. The right plot shows the empirically measured speaker-centric knowledge (or likelihoods), which is the core component of the SCIMPI model.

et al. (2014) used to represent specificity implicatures in their lexical uncertainty model and merged them by taking a simple average and smoothing by adding a small probability to each entry and renormalizing. Using this resultant listener-centric knowledge, I took a Bayesian inversion to get the speaker-centric knowledge for use in the SCIMPI model. Figure 10 shows the SCIMPI model predictions generated using this speaker-centric knowledge and a uniform prior over meanings. With this speaker-centric knowledge, SCIMPI captures the same trend as the lexical uncertainty model, but in a smoother manner and without the need to posit additional latent set-theoretic lexical representations.

Because of SCIMPI's speaker-centric structure, it is unnecessary to start with the set-theoretic lexica from the RSA lexical uncertainty model. Instead, we can show that the SCIMPI model can derive these same implicatures directly using a speaker-centric knowledge distribution. This native or direct approach illustrates the generality of the SCIMPI model, since the same model structure is used to account for a phenomenon that RSA needed auxiliary assumptions to capture. Because experimental data were not available for this speaker-centric knowledge distribution, I generated a hypothetical speaker-centric knowledge distribution by starting with what seem like prima facie reasonable values given my own intuitions. Though such a procedure certainly does not suffice to show the applicability of the model for people's actual speaker-centric knowledge in the case of the "some" vs. "all" specificity implcaiture, it does allow for the illustration of how the model's dynamics might generate such implicatures. The probabilities of saying "All" vs. "Some" to communicate the meaning $\exists\neg\forall$(some but not all) were chosen based on the simple reasoning that a speaker will almost certainly say "Some" here because "All" does not literally apply, but that there is some small probability that she might still say "All." The probabilities that the speaker would say "All" vs. "Some" to communicate the meaning $\forall$(all) were chosen to be equal to illustrate the effect[14]. While these probabilities might be generated by some model, they seem likely to simply arise from knowledge of how people produce language. Figure 11 shows the example values chosen for the production distribution as well as the resultant implicature as generated by the SCIMPI model[15].

[14]If instead the speaker is more likely to say "All" than "Some" in this case but with a lesser discrepancy than in the $\exists\neg\forall$ case, the predictions are simply strengthened.

[15]I am aware of the potential biases introduced by what might pugnaciously be called "making up" the distribution. However, these numbers really are easy to motivate and it is simple to show why we might think of them as reasonable. I have been careful to avoid determining these numbers post-hoc and therefore the numbers shown were the first ones I came up with. It is also possible to show that small perturbations of these numbers do not result in any large changes in the model predictions and that the full range of speaker-centric knowledge here gives rise to predictions that are reasonable given that knowledge. However, due to the difficulty of graphically visualizing the predictions across this full parameter space, I have opted to present a single illustrative example rather than an exhaustive characterization of the model dynamics in this scenario. It is also worth noting that the decisions I have made here are not so different from the decisions that need to be made my RSA model crafters

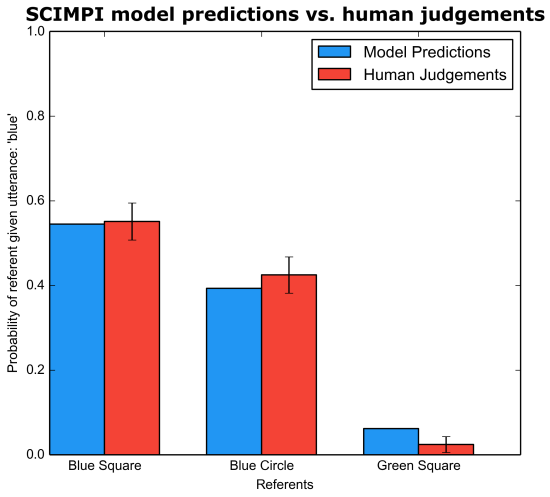

Figure 8: This figure juxtaposes the SCIMPI model predictions about what the speaker is likely to be referring to by producing the utterance "blue" against the participants' corresponding judgments. All error bars show 95% confidence intervals.

Looking at these results, some might think this to be a trivial prediction. Of course these differential speaker preferences are going to produce this differential listener inference. This is true. The dynamics here are not complicated. The lack of complexity here, however, is not a defect in the model's account, but is reflective of the lack of complexity of the inference itself. While models such as RSA make such pragmatic inferences appear to involve arcane and complex computations, SCIMPI obviates the essential simplicity involved. Specificity implicatures–like their ad hoc cousins–can be modeled solely using Bayesian counterfactual inference over probabilistic speaker-centric knowledge. It is also worth noting that the predictions illustrated in figure 11 are derived without the need to represent a "null utterance" whose cognitive existence and normative computational relevance are dubious at best.

Relevance effects and expanding the language

Earlier we saw that RSA, while able to represent differential relevance of meanings given a speaker or a context, was unable to represent this same differential relevance for utterances. This was because of the way in which RSA represents language, which fixes the probability of a meaning given an utterance based on the "literal meanings" of the utterances. The lexical uncertainty model is able to avoid this constraint, but like all versions of lexical uncertainty we have seen before, this amelioration comes at the cost of either positing the representation of a super-exponential number of lexica in the number of utterances and meanings, or of imposing rigid limits to the kinds of uncertainty that may be represented.

SCIMPI, however, is able to represent differential relevance of utterances in a natural way without requiring any auxiliary structures. In SCIMPI, the relevance of an utterance is simply given by the probability of generating that utterance independent of the intended meaning. This may be formally computed as $relevance(u) = p(u|c) = \sum_{m \in M} P(u|m,c)p(m|c)$. If an utterance is unlikely to be generated to communicate each of the meanings that are being considered (each of the relevant meanings), then it will play a minimal role in counterfactual inferences, and its presence or absence in the alternative utterance set will have little effect on the overall model predictions. Importantly, this contextual irrelevance does not prevent the utterance from having strong specificity. It might be that the utterance "demitasse" is one hundred times as likely to be produced in order to communicate about a cup than to communicate about a table (i.e.

when designing a lexicon or a set of candidate lexica. In the future, it should be possible to directly glean this speaker-centric knowledge distribution from human participants using a carefully designed experiment.

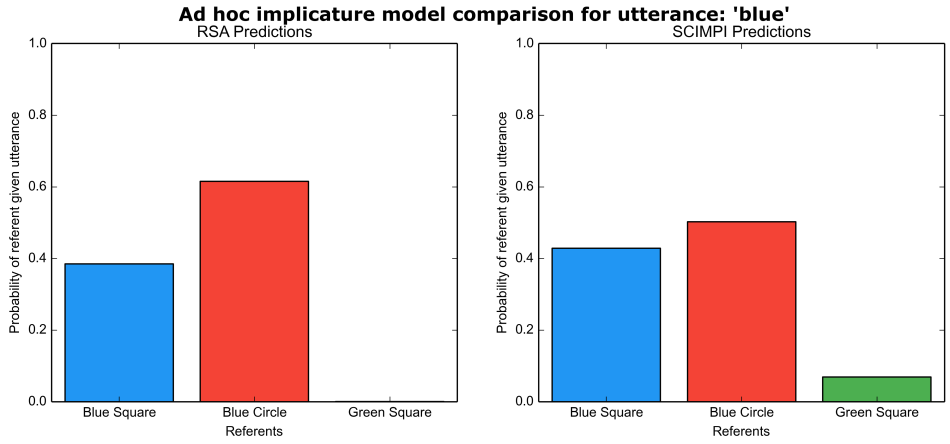

Figure 9: This figure shows the model predictions for RSA and SCIMPI using the empirical priors from the Frank and Goodman (2012) RSA study. Note how the SCIMPI model captures the same general trend as the RSA model, but also allows for the small probability that the speaker might use "blue" to refer to the green square.

$p(\text{“}dematisse\text{”}|cup) = 100 * p(\text{“}dematisse\text{”}|table)$), but it may still be fifty times less likely to be produced in order to communicate about the cup than the alternative utterance "cup" (i.e. $p(\text{“}cup\text{”}|cup) = 50 * p(\text{“}dematisse\text{”}|cup)$). This ability to represent independence between utterance informativity and utterance relevance allows SCIMPI to represent a natural language with millions of useful utterances, but where only a few relevant utterances are being considered at a time. In addition, this property allows SCIMPI to represent the strong implicatures that would be generated if a speaker were to produce an utterance that the listener thought to be irrelevant. While we will not provide a formal illustration of such a scenario, it is clear that this would provide a strong counterfactual trade-off (a strong implicature) and would allow the listener to infer a lot of information about what the speaker might have meant by this utterance. This inferential strength is likely used in practice to convey auxiliary meaning information (other than the literal referent), which may be represented in a more complex multidimensional meaning space such as the reference-affect space modeled by Kao et al. (2014b).) SCIMPI deliberately does not provide a direct mechanism by which to compute the relevance of an utterance given a scenario (it is built into the speaker-centric knowledge) as a design feature, allowing the model to remain maximally independent of specific theories of relevance and salience while retaining the full ability to represent and perform pragmatic computation on knowledge provided by such theories.

In order to illustrate SCIMPI's ability to produce implicatures that are invariant to a growing set of contextually irrelevant or unlikely alternative utterances, I have ran a similar simulation to the equivalent RSA case shown in figure 5. To do this, I included the same sets of meanings and alternative utterances as in the RSA case, but took advantage of SCIMPI's ability to represent contextual irrelevance by lowering the overall probability of the unlikely additional utterances "Regular Quadrilateral" and "Bluesquare". Both of these utterances are still significantly more likely to refer to referents that satisfy their "literal" denotation (the set of squares for "Regular Quadrilateral" and the blue square for "Bluesquare") and therefore retain their specificity, but their overall probability of being produced is very low. Because of this, we expect to see minimal change in the posterior model predictions, since the listener does not believe that the speaker is seriously considering producing these utterances. Figure 12 illustrates the SCIMPI model predictions for this scenario.

As you can see, SCIMPI is capable of representing how an expanded but contextually irrelevant utterance set would not significantly alter model predictions. Unlike RSA, this makes it a candidate for representing the pragmatic effects of differential relevance as predicted by informal relevance theoretic accounts. Of course, pragmatic inferences are increasingly affected by increasingly relevant alternative utterances (ex: perhaps you have reason to believe that the speaker likes to say "Regular Quadrilateral" or "Bluesquare"), and the continuous representation of uncertainty built into the SCIMPI model is able to smoothly account for these

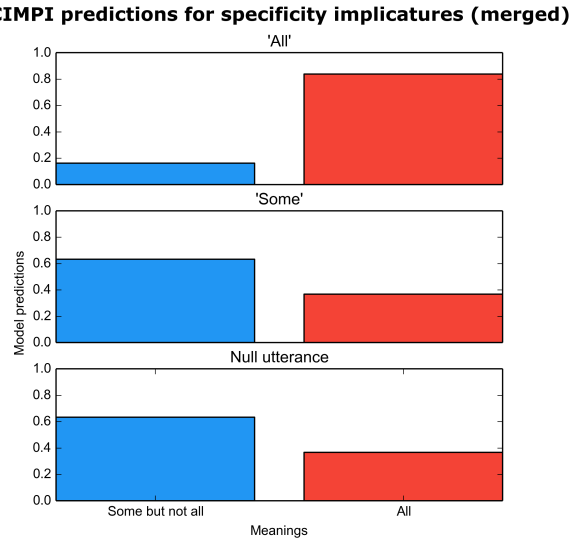

Figure 10: SCIMPI predictions for the "Some" vs. "All" specificity implicature with the speaker-centric knowledge generated by merging the lexica from the lexical uncertainty RSA model of this scenario.

variations. Though the lexical uncertainty variant of RSA is in theory capable of producing these predictions given some set of parameters, it does a less than ideal job of it. First of all, the lexical uncertainty model computes pragmatic inferences by performing counterfactual inference on single lexica and then merging the results. This means that a model that contains separate lexica with the contextually irrelevant utterances would be capable of representing the low prior probability of these utterance, but not the pragmatic inferences generated by these utterances. If these irrelevant utterances were not produced, then these predictions might be reasonable, but if they were produced, they would not be capable of driving the strong implicatures we would expect because they are not counterfactually "competing" with more relevant utterances. In other words, the lexical uncertainty model does not capture the inference that the speaker chose that alternative utterance for a reason and that this conveys information. Instead, it is not even clear how the lexical uncertainty model would handle this scenario, since the predictions lexica with a high prior probability that do not contain this utterance would not be able to computed at all given that they cannot represent the fact that an utterance outside their purview was produced. Additionally, the lexical uncertainty model would need to represent this differential relevance, not in terms of the relevance of utterances, but in terms of the "relevance" of lexica. This means that the model would be making the implicit theoretical claim that people use entirely independent languages in different scenarios. Though I do not have empirical evidence against this, it does seem prima facie implausible given the lack of evidence for this kind of representation throughout the psycholinguistics and neurolinguistics literatures. Finally, the lexical uncertainty model is not ideal for this case because of the aforementioned scaling problems. SCIMPI captures the pragmatic inferences we expect from realistic interlocutors with large lexica and meaning sets who are situated in a particular context, and is able to represent the predictions provided by contemporary relevance theoretic accounts of pragmatic inference.

Discussion

Models and Approaches

So far we have articulated two alternative accounts of pragmatic comprehension–the RSA family of models, and the SCIMPI model–and compared their performance on three empirical phenomena central to pragmatic interpretation. Though this has yielded insight as to how these different approaches account for pragmatic comprehension, we have not yet analyzed the full scope of advantages and disadvantages of each approach. In this section we will perform this analysis, both by citing results from the previous sections and by articulating

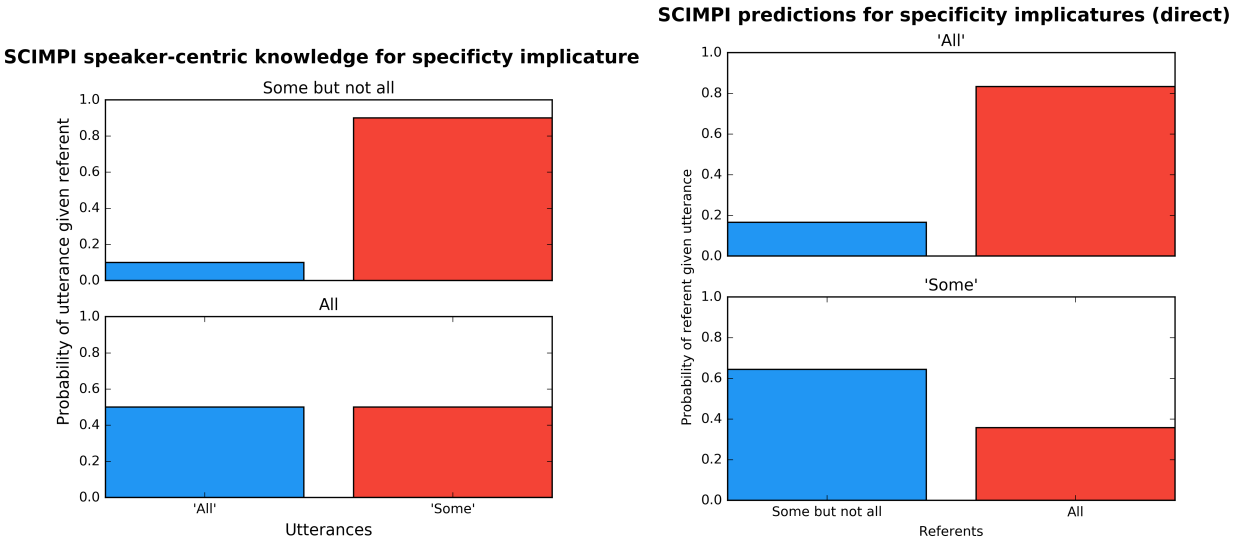

Figure 11: The lefthand plot shows the speaker-centric knowledge used in this "direct" version of SCIMPI for the "Some" vs. "All" specificity implicature. The righthand plot shows the SCIMPI model predictions generated using this knowledge distribution.

theoretical points not yet discussed in sufficient detail. The discussion section is structured into a series of comparisons, each of which is under its own section heading. The four criteria on which the modes will be evaluated are simplicity and parsimony, cognitive plausibility, overtness and scrutabillity of assumptions, and empirical consistency.

Simplicity and parsimony

All other things equal, a simpler model is a better model. It contains fewer unmotivated assumptions and is more likely to generalize beyond the data. Sir Isaac Newton put this quite elegantly when he said "We are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances. Therefore, to the same natural effects we must, as far as possible, assign the same causes (Newton, 1728)." In this vein, the SCIMPI model wins out over RSA. SCIMPI is a truly minimal model in that it includes only the structures absolutely necessary to give rise to the set of phenomena we have labeled as pragmatic inference. All additional structure for handling special cases lies outside of this core model, and belongs within specific theories of these special cases. RSA, on the other hand, includes a number of extra structures representing infinite counterfactual recursion, a highly structured language model, a decision theoretic softmax equation, and multiple possible lexica to be summed over. These structures allow it to represent certain special cases and change the parameterization, but they do not contribute to the model's ability to account for the phenomena. SCIMPI's simpler model structure gives rise to predictions that are at least as good as RSA's, but without the need for the auxiliary structures and causes built into RSA.

In similar vein to Newton's quote on simplicity and parsimony, Bertrand Russell argued for a parsimonious approach by stating, "whenever possible, substitute constructions out of known entities for inferences to unknown entities (Russell, 1924)." Such parsimony allows for the interconnection between theories and for an explanation of the world that does not arbitrarily isolate one phenomenon from another. SCIMPI is strongly motivated by this desire to understand pragmatic interpretation, not separately on a case by case basis, and not as a normative phenomenon independent of cognition, but as a general framework for inference situated in the larger scope of cognition and culture. In this way, SCIMPI makes minimal assumptions about the structure of language representation and about the nature of the process by which a speaker chooses what to say. It externalizes all of these concerns, leaving them to theories of the specific and intertwined cognitive processes involved, and instead posits only one simple computation about how we transform this speaker-centric knowledge into predictions about what people mean by what they say. Because of this modularity,

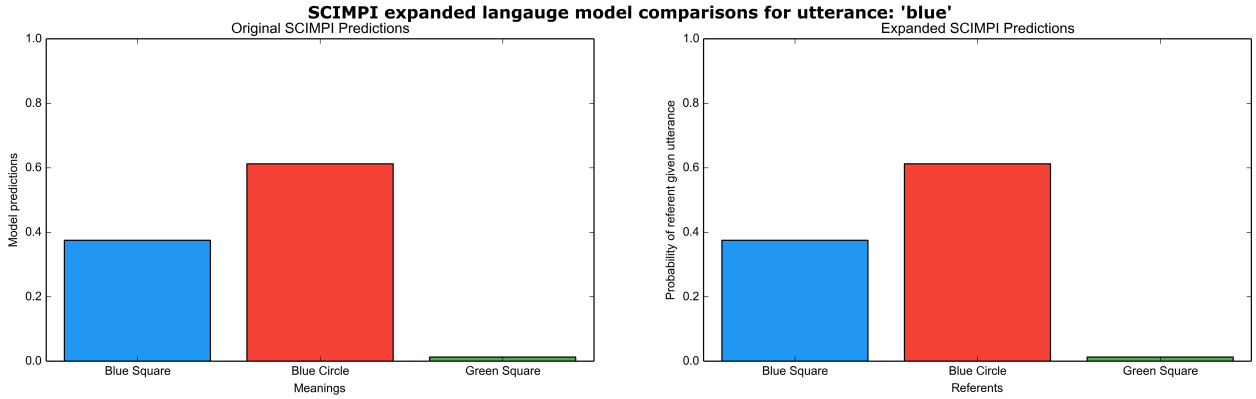

Figure 12: This figure shows the effects of expanding the SCIMPI model with additional utterances that are contextually unlikely. The plot on the left shows the SCIMPI model predictions for the original utterance set. The plot on the right shows the SCIMPI model predictions for the expanded utterance set. The predictions are indistinguishable to the naked eye.

SCIMPI is consistent with a large number of interesting and complex phenomena observed in the cognitive sciences of language and communication. Each of the parameters in SCIMPI has a direct correspondence to a directly measurable quantity, and no additional entities are posited.

RSA, however, includes latent structures such as lexica and a recursive set of interlocutor models for whose existence we do not have evidence. These additional entities would be useful if they did useful work above and beyond the SCIMPI model, but since these models either make inferior predictions (in the case of basic RSA and truncated lexical uncertainty) or do not add anything beyond SCIMPI (in the case of full lexical uncertainty), the scientific pressure towards parsimony suggests that SCIMPI makes for a better theory.

Cognitive plausibility

Ultimately, any successful model of a cognitive process must be implementable in the brain. Though computational level models are specified in terms of abstract functional relationships rather than with algorithmic and implementational detail, the continued improvement and specification of a model will eventually connect with these lower level theories. While RSA and SCIMPI are both at far too early a stage to worry about algorithmic and implementational details, it would behoove the proponents of these models to consider whether such details might plausibly be specified in the future. We have seen that RSA, in order to account for a number of empirically observed phenomena, must include additional structures, transforming it into variants such as the lexical uncertainty model. We have also seen that this variant is only able to represent realistic uncertainty structures if it includes a representation of all or nearly all possible lexica for the sets of possible meanings and utterances. The number of lexica required, however, grows super-exponentially with the numbers of meanings and utterances, and therefore the representation of a sufficient number of lexica to model realistic meaning and utterance sets is intractable, even for the impressive neuronal computers in our heads. Even if we were to somehow find a compatible representational scheme for this set of lexica (far more lexica than the number of particles in the observable universe), the computations required to make even the simplest inferences would be similarly intractable. SCIMPI on the other hand requires the representation of only a single language distribution with a number of entries equal to the product of the number of possible meanings and the number of possible utterances. While this is still a large number, it is well within the range of cognitive plausibility, and is likely to be made even smaller through online computation using of a subset of the distribution (the most relevant subset), which could be easily truncated (this occurs naturally in neural computations) so that only the most relevant meanings and utterances are even considered. Lexical uncertainty, on the other hand, can only be truncated using a much more complicated set of operations that go through the set of possible lexica and remove those which are not considered relevant to the current computations. Furthermore, SCIMPI does not claim that all of the speaker-centric knowledge is stored in

such a distribution, but that this distribution is generated using auxiliary theories of language, meaning, and context. The set-theoretic language mappings in RSA models do not lend themselves as well to this sort of online generation, and they are more or less explicitly claimed to be a stored part of language knowledge. Though it is difficult to perform a thorough cognitive plausibility analysis for computational level models, this heuristic analysis suggests that the SCIMPI model is much more likely to correspond to and be consistent with the cognitive computations performed by human brains than the RSA lexical uncertainty model is.

Overtness and scrutability of assumptions

Another advantage of SCIMPI over RSA comes from the clarity of the model specification and how easy it is to tell which assumptions the model is making. All of the assumptions in SCIMPI are overt and may therefore be challenged or defended directly. RSA, however, makes a number of covert assumptions, which must be extracted by digging into the specifics of the model specification and all of the theoretical claims that it supervenes on. It is not, for example, obvious at first glance that the RSA model assumes that the language being modeled is a strict set of mappings between meanings and utterances and that this entails the model's incapability of representing uncertainty about the meanings of words. Instead, this insight may be extracted only by noticing that no operations present in the RSA model specification are able to transform the zeros in the lexicon entries into anything other than zeros. Because of this, no amount of contextual or pragmatic evidence is ever able to cause RSA to predict reference to entities by utterances that did not contain these referents in their literal meaning sets. Furthermore, RSA makes a number of very strong implicit assumptions about the relationship between the beliefs of the listener and those of the speaker. The decision theoretic and economic principles from which RSA is derived assume that the listener and the speaker share the EXACT SAME beliefs about the nature of the language and that the listener's model of the speaker is known for certain to be true. Any deviation from these assumptions does not only require the representation of uncertainty within the model, but also necessitates a strong reformulation of the model that includes uncertainty about each parameter and each recursive step as well as the addition of separate distributions to represent the uncertainty in the listener's beliefs about the speaker's production and about the listener's beliefs about the speaker's beliefs about the listener's comprehension. These assumptions are prima facie invisible in the RSA model, as they are hidden by the apparent elegance of the Schelling-style recursive inference in which the need to represent all of these extra distributions vanishes only under extreme certainty. To ameliorate this, researchers developed the lexical uncertainty model to allow for uncertainty about the lexica. Far from being an elegant generalization of the RSA model, it simply acts as a field bandage, adding uncertainty about which RSA model is being used, while retaining the inability to represent the full extent of uncertainty present in a probabilistic Schelling model. The lexical uncertainty model continues to require the assumption of perfect mutual knowledge within the pragmatic computations on each lexicon, and cannot represent uncertainty about the number of levels of recursive inference performed, nor about the pragmatic reasoning process in general. While the lexical uncertainty model is, in principle, able to produce any distribution of posterior predictions, it does so only by incorporating massive representations and by requiring combinations of parameters that in general have no theoretical or empirical correspondence to the phenomenon being modeled. SCIMPI, on the other hand, represents listener beliefs directly, requires no additional normative constraints beyond those required for basic Bayesian cognitive models, and is able to accurately predict listener inferences without making any ad hoc assumptions about the structure and specifics of the speaker's production model.

Empirical consistency

Finally, the most commonly analyzed desideratum of a scientific theory is its consistency with the data. Much of this was shown earlier in the paper in the analyses on empirical cases, but there are other data that we can compare each model to, albeit in a less formal manner. First of all, we know that language is learned. Though there are human biases that affect our learning, the prevalence of different languages with vastly different syntactic structures, meanings, and utterance-meaning mappings, suggests that a large component of linguistic structure is acquired through experience. This means that any plausible model of language must be learnable. This argument was made by Chomsky (1957) to support his generative grammatical theory, and while the details of such an argument are vigorously debated, almost everyone agrees on the general

learnability requirement. One of the key problems with RSA is that its set theoretic language representations cannot be plausibly acquired. As humans are never privy to an infallible observation of the relationship between meanings and language, a representation of language that does not include uncertainty cannot be incrementally acquired through uncertain experience. While it is possible to learn such a representation given a structured hypothesis space of possible languages and continuous weights over these languages a la the lexical uncertainty model, such a framework for acquisition would require that all possible languages be represented and that a very complex process of combinatorial latent inference over language weights be performed every single time the agent's language model needs to be updated. Because of these limitations, it is safe to say that, without additional theory connecting the language representation in the pragmatic models to a native language representation containing uncertainty, the RSA language model is unlearnable and the lexical uncertainty language model cannot be tractably acquired or represented. SCIMPI's language model, on the other hand, is directly acquirable through the kinds of observations and statistical inferences that we make everyday, and which a number of contemporary language learning theories suggest comprise our linguistic acquisition processes (Xu and Tenenbaum, 2007; Yu and Smith, 2007; Yu et al., 2007; Smith and Yu, 2008; Frank et al., 2008). Incremental acquisition of SCIMPI's speaker centric knowledge (at least for the case of minimally-structured acquisition processes like word learning) requires only a prior distribution capturing human cognitive biases as they apply to language acquisition and observations of language being used by other agents. Because of the naturalness and tractability with which it's representation of language knowledge affords acquisition by humans in the world, SCIMPI is much more consistent with language acquisition than RSA is.

SCIMPI also fits more neatly with all of the rich social and cognitive phenomena that have been shown by empirical studies on language use and language processing. These include metaphor, alignment, deception, and more(Lakoff and Johnson, 2008; Pickering and Garrod, 2006). While the model does not explicitly capture these phenomena, it's minimality allows it to remain consistent with them and therefore to serve as the basis for a unified model of all of these properties of human language and cognition and how they fit together to yield the way people use language in the world.

Potential impacts of SCIMPI on linguistic theory

With a simple parsimonious model of pragmatic comprehension, linguistic pragmatics and other related fields can begin to make new progress on a number of outstanding problems. Most directly, SCIMPI enables us to provide a unified account of many pragmatic phenomena discussed in the literature, demonstrating that they are not as disparate as is currently thought. Ad hoc, specificity, and ignorance implicatures, for instance, are not directly distinguishable in a SCIMPI account of pragmatic interpretation. Each of them is comprised solely of a Bayesian counterfactual inference given a listener's current expectations about the speaker's production. While delineation of these phenomena may still be useful for running controlled experimental studies, the availability of a unified account allows us to clearly see their commonalities alongside their differences.

In addition, if we notice the smooth connection between the pragmatic predictions of SCIMPI and the acquisition of new lexical and linguistic knowledge through experience, we can better appreciate the theoretical continuity between language acquisition and language use. Though these fields are traditionally separated, there is lots of reason to believe that language acquisition continues throughout our experience. Given this life-long acquisition and modification of our linguistic knowledge, we would like an account of language use and acquisition that blurs the traditional lines and demonstrates this continuity. Though this would require some additional model structure to be added to the SCIMPI framework, it is not too difficult to see how such an account might work. Since SCIMPI is a Bayesian model, and since there are a number of Bayesian accounts of language acquisition, the merging of such accounts would be a fruitful endeavor for exploring this relationship between acquisition and use.

Besides the continuity between acquisition and use, SCIMPI also makes room for the possibility of a continuous spectrum between word-sense disambiguation and pragmatic inference. Word-sense disambiguation refers to the process by which a listener determines which "sense" of a word is intended when a polysemous (a word with multiple meanings or senses) word is produced. For example, if I see the word "bass" in print, I need to use contextual factors to determine whether it refers to the fish or to the acoustic frequency range. Though this case involves well-differentiated alternative meanings, some words are often used in

different ways that do not clearly constitute distinct senses. For example, adjectives such as "wonderful" are frequently both "literally" or sarcastically, which constitute alternative meanings. This distinction can be conceptualized either in terms of multiple senses being associated with the word or as the application of some sort of sarcasm function to the word in order to invert its meaning. Due to this theoretical duality and the continuity with which it bleeds into clear cut cases of classical pragmatic inference and traditional word-sense disambiguation, we would like to be able to represent all of these phenomena in a continuous manner whereby a single model is capable of generating all of the phenomena, and where the differences in phenomena are captured by differences in parameter values or auxiliary model structures. The general nature of SCIMPI as a model of pragmatic inference allows it to do just this. SCIMPI is capable of capturing and representing all of these phenomena and therefore allows us to see the continuity that exists between word-sense disambiguation and pragmatic inference, despite the traditional separation between them. An additional advantage of this theoretical move is the ability to inspire general purpose learning algorithms for pragmatic inference. Instead of treating word-sense disambiguation and pragmatic inference as independent operations, we can attempt to learn the intricacies of both in terms of a single continuous (though not necessarily smooth and linear) representation. Of course we do not yet know how to build such an algorithm, but the rate of progress in natural language processing and the theoretical realization of this continuity suggest that it might be possible in the near-ish future.

Another impact of SCIMPI is its ability to provide a formal account of pragmatic inference that is consistent with a large amount of contemporary cognitive scientific and linguistic work that has only so far been demonstrated in an informal manner. In addition to the case of relevance theory demonstrated above, SCIMPI is consistent with the theories of common ground(Clark and Marshall, 1981), conversational alignment(Pickering and Garrod, 2006), and interactive inference in discourse(Clark and Wilkes-Gibbs, 1986; Sacks et al., 1974). It is also more flexible than Gricean maxims and Horn's principles(Grice, 1975; Horn, 1984), and therefore allows for the smooth incorporation of human communicative features such as deception and irrationality, which are not easily accounted for in these traditional frameworks. This cognitive consistency provides a potential bridging point between the formalist traditions in cognitive science and linguistics, which traditionally sacrifice generality for precision, and the informalists who provide a compelling picture of cognition as a whole, but have difficulty providing precise predictions. Though SCIMPI does not in itself include a formal account of the complexities of cognition, it offers an approach with which one might be slowly developed through the independent construction of modular theories of small phenomena and the fitting together of these pieces into a complete puzzle.

Conclusion

The introduction of computational models of pragmatics–especially the Bayesian rational speech act model–has made room for new progress in understanding pragmatic inference, as well as for grounding precise formal theories directly in data. Though RSA captures a number of diverse phenomena, we have seen that its accounts are neither as simple nor as parsimonious as they could be. The core RSA model has difficulty representing the intricacies and uncertainties of realistic language use and requires a number of modifications and additions to capture phenomena beyond simple ad hoc implicatures. One such extension, the lexical uncertainty model, initially looked promising due to its ability to represent uncertainty about the language itself. While this was a natural step towards a model that takes epistemic uncertainty about the interlocutor more seriously, it turned out to have a number of problems of its own, most of which derived from the implicit assumptions of the base RSA model. In order to more cleanly and generally represent this uncertainty, we had to take a step back from the RSA specification and reevaluate its assumptions. Most notably, this effort eliminated the assumptions that language is comprehended through recursive interlocutor inference and that language is represented as a set of discrete mappings between meanings and utterances. The resultant model–SCIMPI–was able to capture all of the phenomena we explored with RSA, as well as a number of others that RSA had more trouble with. The theoretical modularity of SCIMPI means that it is able to capture the core component of pragmatic inference–how listeners go from beliefs about how the speaker tends to use to language to inferences about what the speaker meant when she chose to produce a specific utterance–without theoretically impinging upon other related processes. This means that SCIMPI while minimal, is able to make precise predictions about a number of implicature-type inferences, which have

been at the center of the study of pragmatics for decades, while remaining consistent with the cognitive and communicative complexities that pervade real language use. This insight that pragmatic interpretation is fruitfully understood as a combination of simple Bayesian inference with more complex inferences about the speaker's production probabilities may serve as a foundation for the continued exploration of pragmatics by proposing and testing simple and parsimonious theories of the many diverse phenomena that comprise how we as humans, immersed in immense uncertainty about each others' minds and the world, are able to communicate so successfully.

Acknowledgements

I would like to thank my advisory committee of Jeff Elman, David Kirsh, and Ben Bergen for all of their advice, feedback and support while I pursued this unconventional second year project. I would also like to thank Andy Kehler, Dave Barner, Herb Clark, and Marta Kutas for the inspiration and support they provided without any obligation to do so. In general the UCSD Department of Cognitive Science and especially my second year cohort served as a family throughout this endeavor and I really doubt I would have made it this far without them. Finally, I would like to thank my friends and family; while I did not always appear to appreciate their advice and sanity-providing presences, they also were essential to this effort.

It is also important to me to reiterate my indebtedness to Mike Frank, Noah Goodman, and others for their inspiring work on the RSA model. Without this simple yet powerful idea that pragmatic inference could be structured as a Bayesian computation, I would have had nowhere to begin and would not likely have entered the field of pragmatics at all.

References

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

Bergen, L., Levy, R., and Goodman, N. D. (2014). Pragmatic reasoning through semantic inference.

Chierchia, G., Fox, D., and Spector, B. (2008). The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. *Unpublished manuscript*.

Chomsky, N. (1957). *Syntactic Structures*. Mouton, The Hague.

Clark, H. H. and Marshall, C. R. (1981). Definite reference and mutual knowledge.

Clark, H. H. and Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, 22(1):1–39.

Degen, J., Tessler, M. H., and Goodman, N. D. (2015). Wonky worlds: Listeners revise world knowledge when utterances are odd.

Frank, M. C. and Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336.

Frank, M. C. and Goodman, N. D. (2014). Inferring word meanings by assuming that speakers are informative. *Cognitive psychology*, 75:80–96.

Frank, M. C., Goodman, N. D., and Tenenbaum, J. B. (2008). A Bayesian framework for cross-situational word learning. In *Advances in Neural Information Processing Systems 20*.

Franke, M. (2013). Game theoretic pragmatics. *Philosophy Compass*, 8(3):269–284.

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

Grice, H. (1975). Logic and conversation. *Syntax and Semantics*, 3:41–58.

Griffiths, T. L., Lieder, F., and Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in cognitive science*, 7(2):217–229.

Horn, L. (1984). Toward a new taxonomy for pragmatic inference: Q-based and r-based implicature. *Meaning, form, and use in context: Linguistic applications*, pages 11–42.

Hsu, A. and Griffiths, T. L. (2009). Differential use of implicit negative evidence in generative and discriminative language learning. In *Advances in neural information processing systems*, pages 754–762.

Jäger, G. (2008). Game theory in semantics and pragmatics. *manuscript, University of Bielefeld*.

Kao, J. T., Bergen, L., and Goodman, N. D. (2014a). Formalizing the pragmatics of metaphor understanding. In *Proceedings of the 36th annual meeting of the Cognitive Science Society*, pages 719–724.

Kao, J. T., Levy, R., and Goodman, N. D. (2013). The funny thing about incongruity: A computational model of humor in puns. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, pages 728–733.

Kao, J. T., Wu, J. Y., Bergen, L., and Goodman, N. D. (2014b). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, 111(33):12002–12007.

Keynes, J. M. (2006). *General theory of employment, interest and money*. Atlantic Publishers & Dist.

Lakoff, G. and Johnson, M. (2008). *Metaphors we live by*. University of Chicago press.

Lane, L. W. and Ferreira, V. S. (2008). Speaker-external versus speaker-internal forces on utterance form: Do cognitive demands override threats to referential success? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(6):1466.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Henry Holt and Co., New York, NY.

Mollica, F., Piantadosi, S. T., and Tanenhaus, M. K. (2015). The perceptual foundation of linguistic context. *Cognitive Science*.

Muhlstein, L., Frank, M. C., Potts, C., and Levy, R. (2015). Pragmatic coordination on context via definite reference. In *Experimental Pragmatics*.

Newton, I. (1728). Philosophiæ naturalis principia mathematica.

Pickering, M. J. and Garrod, S. (2006). Alignment as the basis for successful communication. *Research on Language and Computation*, 4(2-3):203–228.

Russell, B. (1924). Logical atomism.

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *language*, pages 696–735.

Schelling, T. C. (1960). The strategy of conflict. *Cambridge, Mass*.

Smith, L. and Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3):1558–1568.

Sperber, D. and Wilson, D. (1986). *Relevance: Communication and Cognition*. Blackwell Publishers, Oxford, UK.

Tenenbaum, J. and Griffiths, T. (2002). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, 24:629–640.

Tenenbaum, J. B. (1999). *A Bayesian framework for concept learning*. PhD thesis, Citeseer.

Xu, F. and Tenenbaum, J. (2007). Word Learning as Bayesian Inference. *Psychological Review*, 114:245.

Yu, C. and Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, 18(5):414–420.

Yu, C., Smith, L., Klein, K., and Shiffrin, R. (2007). Hypothesis testing and associative learning in cross-situational word learning: Are they one and the same. In *Proceedings of the 29th Annual Meeting of Cognitive Science Society (CogSci 2007), Nashville, TN*. Citeseer.