A practical introduction to the Rational Speech Act modeling framework

Gregory Scontras
University of California, Irvine

Michael Henry Tessler
Massachusetts Institute of Technology

Michael Franke
University of Osnabrück

Abstract Recent advances in computational cognitive science (i.e., simulation-based probabilistic programs) have paved the way for significant progress in formal, implementable models of pragmatics. Rather than describing a pragmatic reasoning process in prose, these models formalize and implement one, deriving both qualitative and quantitative predictions of human behavior—predictions that consistently prove correct, demonstrating the viability and value of the framework. The current paper provides a practical introduction to and critical assessment of the Bayesian Rational Speech Act modeling framework, unpacking theoretical foundations, exploring technological innovations, and drawing connections to issues beyond current applications.

Keywords: Rational Speech Act models, probabilistic pragmatics, Gricean reasoning

1 Introduction

Much work in formal, compositional semantics follows the tradition of positing systematic but inflexible theories of meaning. In practice, however, the meanings listeners derive from language are heavily dependent on nearly all aspects of context, both linguistic and situational. To formally explain these nuanced aspects of meaning and better understand the compositional mechanism that delivers them, recent work in formal pragmatics recognizes semantics not as one of the final steps in meaning calculation, but rather as one of the first. Within the Bayesian Rational Speech Act (RSA) framework (Goodman & Frank 2016, Franke & Jäger 2016), speakers and listeners reason about each other's reasoning about the literal interpretation of utterances. The resulting interpretation necessarily depends on the literal interpretation of an utterance, but is not necessarily wholly determined by it. This move—reasoning about likely interpretations—provides ready explanations for

complex phenomena ranging from metaphor (Kao et al. 2014a) and hyperbole (Kao et al. 2014b) to the specification of thresholds in degree semantics (Lassiter & Goodman 2013).

The probabilistic pragmatics approach leverages the tools of structured probabilistic models formalized in a stochastic λ-calculus to develop and refine a general theory of communication. The framework synthesizes the knowledge and approaches from diverse areas—formal semantics, Bayesian models of reasoning under uncertainty, formal theories of measurement, philosophy of language, etc.—into an articulated theory of language in practice. These new tools yield improved empirical coverage and richer explanations for linguistic phenomena through the recognition of language as a means of communication, not merely a vacuum-sealed formal system. By subjecting the heretofore off-limits land of pragmatics to articulated formal models, the rapidly-growing body of research both informs pragmatic phenomena and enriches theories of linguistic meaning.

These models are particularly well-suited for capturing complex patterns of pragmatic reasoning, especially cases that integrate multiple sources of uncertainty. For example, in order to infer what a speaker has meant by an utterance in a given context, listeners may have to reason about lexical or syntactic ambiguity (Bergen et al. 2016, Savinelli et al. 2017, Franke & Bergen 2020), resolve underspecification in semantic meaning (Lassiter & Goodman 2013), or consider the possibility of non-literal interpretation (Kao et al. 2014a,b). Listeners may also have to reason about the speaker's own likely epistemic state (Goodman & Stuhlmüller 2013, Scontras & Goodman 2017, Herbstritt & Franke 2019) or motives the speaker might entertain other than pure information exchange (Yoon et al. 2016, 2017). Probability calculus provides a well-understood method for formalizing such complex patterns of reasoning, grounded in norms of rational belief formation and action choice. Recent advances in computer science provide accessible computational methods for implementing complex models of probabilistic reasoning, enabling us to naturally formulate and conveniently test pragmatic theories at a level of complexity and formal rigor that far exceeds what was possible before.

The current paper offers a practical introduction to the modeling framework, serving as a companion piece to the hands-on web-book at www.problang.org (Scontras et al. electronic). We begin in Section 2 with a high-level overview of RSA, walking through its basic implementation and the philosophical foundations that informed the architectural choices. We then explore variations to the basic architecture in Section 3, surveying technological innovations that have allowed for broader empirical coverage. After a brief excursus on the

relative benefits of "thinking in math" vs. "thinking in code" in Section 4, in Section 5 we discuss common practical considerations facing the modeler. In Section 6, we explore limitations of the current framework, which also serve as guidance for future extensions. Section 7 concludes.

2 High-level overview of RSA

The RSA framework views language understanding as a process of recursive social reasoning between speakers and listeners: listeners interpret the utterances they hear by reasoning about how speakers generate them; speakers choose their utterances by reasoning about how listeners interpret them. In the basic RSA model from Frank & Goodman (2012) (hereafter, "Vanilla RSA"), this recursion involves three layers of inference.[1] Typically formulated as statements of conditional probability, as in (1), (3), and (4), these inference layers correspond to models of speakers and listeners. We will go over these definitions step by step.

The reasoning grounds out in the naive, literal listener, L_0, who interprets utterances according to their literal semantics.

$$P_{L_0}(s \mid u) \propto P_{L_0}(s) \cdot [\![u]\!](s) \tag{1}$$

The semantics of utterance u is captured in the meaning function $[\![u]\!] : s \mapsto [0;1]$, which maps states to truth values. We assume binary truth-values here, but the formalism works just as well for non-binary, fuzzy values (e.g., Degen et al. 2020). Given a binary truth-conditional semantics, the rule in Equation (1) is equivalent to the following formulation in terms of belief update with the set of states where u is true:

$$P_{L_0}(s \mid u) = P_{L_0}(s \mid \{s \mid [\![u]\!](s) = 1\}) \tag{2}$$

According to Equation (1), L_0 hears some utterance u, and updates their prior beliefs $P_{L_0}(s)$ about the world states s with the information that u is true. If, as we will assume throughout unless noted otherwise, the prior beliefs of the literal listener are a uniform distribution (i.e., equal probability for all states), (1) returns a uniform probability distribution over the states s that u maps to `true`.[2]

1 For an earlier proposal of this architecture, see Benz & van Rooij (2007).

2 For more discussion on using flat priors for the literal listener, at least in referential-communication games, see Qing & Franke (2015). Some models may wind up using an informative prior at the level of L_0, as in the context-inference models presented in Section 3.3 below.

One layer up, a pragmatic speaker, S_1, chooses utterances in proportion to their utility U_{S_1}:

$$P_{S_1}(u \mid s) \propto \exp(\alpha \cdot U_{S_1}(u;s)), \text{ where} \tag{3}$$
$$U_{S_1}(u;s) = \log P_{L_0}(s \mid u) - C(u)$$

Utterances are useful to the extent that they maximize the probability that L_0 will infer the correct s on the basis of u (i.e., informativity), while minimizing the cost $C(u)$ of u (speakers aim to be efficient). So, when selecting utterances, S_1 considers their effect on interpretation (i.e., on L_0's resulting beliefs); utterances that are most likely to lead L_0 to the correct belief are most likely to be chosen by S_1. The numerical utilities are mapped onto discrete choice probabilities by a soft-max function with parameter α. The larger α is, the more the speaker's choice probabilities converge to a strict maximization of utility.

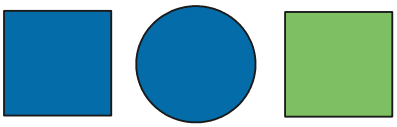
At the top layer of inference, the pragmatic listener, L_1, interprets utterances to update their prior beliefs $P_{L_1}(s)$ by taking into account how likely the speaker would have been to produce the observed utterance u in various states:

$$P_{L_1}(s \mid u) \propto P_{L_1}(s) \cdot \ P_{S_1}(u \mid s) \tag{4}$$

In other words, unlike L_0, who reasons directly about the utterance semantics, L_1 reasons instead about the process that generated the utterance; that process is the speaker S_1. L_1 thus infers s on the basis of u by considering which states are a priori likely and reasoning about the probability that S_1 would have chosen u to signal s to L_0. Because L_1 reasons about S_1, who in turn reasons about the literal semantics in L_0, L_1's interpretation is affected by the semantics of u, albeit only indirectly via the S_1 layer.

To understand the Vanilla RSA model better, it will help to consider a concrete example. In its initial formulation, Frank & Goodman (2012) used the RSA framework to model referent choices in efficient communication. Suppose we are in a context as in Figure 1 (panel A) with three objects: a blue square, a blue circle, and a green square. Suppose further that a speaker is trying to signal a single object in this world to a cooperative listener, and that the speaker can only use a single-word utterance to do so. The utterances available to the speaker are "blue", "green", "square", and "circle"; the possible states the listener might infer correspond to the three objects: blue-square, blue-circle, and green-square. We have the expected truth-functional semantics for the utterances: "blue" maps blue-square and blue-circle to true but green-square to false, "green" maps blue-square and blue-circle to false but green-square to true, etc.

A: context



B: literal listener

			
"blue"	0.5	0.5	0
"green"	0	0	1
"square"	0.5	0	0.5
"circle"	0	1	0

C: speaker

	"blue"	"green"	"square"	"circle"
	0.5	0	0.5	0
	0.33	0	0	0.66
	0	0.66	0.33	0

D: pragmatic listener

			
"blue"	0.6	0.4	0
"green"	0	0	1
"square"	0.6	0	0.4
"circle"	0	1	0

Figure 1 An example of the Vanilla RSA model applied to a referential-communication game from Frank & Goodman (2012). Panel A shows the context of conversation with three potential referents. Panel B shows the interpretation probabilities of the literal listener. Panel C shows the production probabilities predicted for the pragmatic speaker when assuming $\alpha = 1$. Panel D shows the interpretation probabilities of the pragmatic listener, based on the production probabilities in panel C. All calculation assume flat priors, for the literal and the pragmatic listener.

With the semantics as stated, L_0 interprets utterances to return uniform probability distributions over the compatible states (see Figure 1 panel B). Hearing the utterance "blue", L_0 returns a belief distribution over states that divides probability equally between blue-square and blue-circle, the only objects compatible with the semantics of "blue". Hearing "circle", L_0 returns a distribution with 100% of the probability on blue-circle—L_0 is certain that the speaker intends to signal blue-circle. With L_0 formulated in this way, we have an agent who interprets utterances naively, according to the literal semantics.

S_1 chooses utterances by simulating their effect on L_0. Suppose the speaker wants to communicate blue-circle to L_0. Two utterances, "green" and "square", stand no chance of communicating the intended state to L_0 and so

they are ruled out entirely. The other two utterances, "blue" and "circle", are both literally compatible with blue-circle, but one of the utterances is more likely to lead L_0 to the correct belief state. If the speaker were to utter "blue", we saw that L_0's belief distribution would be evenly split between blue-square and blue-circle. In other words, "blue" has a 50% chance of leading L_0 to the correct state. On the other hand, "circle" has a 100% chance of leading L_0 to the correct state. Assuming equal utterance costs, "circle" is thus twice as useful to the speaker as "blue" for communicating blue-circle to L_0, and S_1 reflects this asymmetry in utility by, if we assume $\alpha = 1$, assigning twice as much probability to "circle" in the posterior distribution over utterance choices (see Figure 1 panel C). For communicating the blue-square state, both "blue" and "square" have a 50% chance of leading L_0 to the correct state, so both utterances are equally useful and thus equally probable in the S_1 posterior.

Within this simple reference-game scenario, L_1 reasons pragmatically about S_1 to break the symmetry in the semantics (see Figure 1 panel D). Hearing "blue", L_1 will invert the S_1 model to determine which state (i.e., which object) the speaker is most likely trying to communicate. Had the speaker wanted to communicate the blue-circle state, we saw above that the speaker would have been more likely to utter "circle". But the speaker did not utter "circle"; she uttered "blue" instead. This counterfactual reasoning leads L_1 to down-weight the probability of the state that "circle" would have uniquely picked out, since the speaker could have said "circle" but chose not to. As a result, more probability gets assigned to the blue-square state, the other state compatible with the semantics of the utterance. In this way, we capture the Gricean specificity implicature associated with uttering "blue" in a scenario as in Figure 1: the speaker probably intends the blue square because if she wanted to communicate the blue circle she could have said "circle".

Prior beliefs & pragmatic content A key component of the RSA framework is the updating of prior beliefs: listeners use the utterances they hear as signal with which to update their beliefs about the world. In walking through the reference-game example above, we assumed a uniform prior over world states: before hearing the speaker's utterance, L_1 had no reason to suspect that any object was more or less likely to get referenced. With a different, more informative prior, it is possible to shift the qualitative predictions of L_1. Suppose that L_1 has reason to suspect that the blue-circle state is most likely to get referenced. Now there are two opposing pressures operating on L_1's interpretation of the utterance "blue" (i.e., on L_1's posterior beliefs):

the prior belief favoring blue-circle over blue-square, and the specificity inference discussed above that favors blue-square over blue-circle (if the speaker had wanted to reference blue-circle, she could have said "circle"). With a sufficiently strong prior in favor of blue-circle, it is possible that the most likely interpretation of "blue", according to an *a priori* biased L_1, is blue-circle after all.

To dissect the impact of prior beliefs and genuine pragmatic interpretation at the level of the pragmatic listener, we must therefore compare the pragmatic listener's beliefs under a literal interpretation $P_{L_1}(\cdot \mid \{s \mid [\![u]\!](s) = 1\})$ against the pragmatic listener's beliefs under a pragmatic interpretation $P_{L_1}(\cdot \mid u)$, both of which are probability distributions over states. Following Skyrms (2010), we might look at the information about each state s carried by the pragmatic interpretation of u on top of its semantic interpretation, like so:

$$\text{Info}(s,u) = P_{L_1}(s \mid u) - P_{L_1}(s \mid \{s \mid [\![u]\!](s) = 1\})$$

Dropping quantitative aspects from the picture, we may then say that *utterance u pragmatically conveys s* iff $\text{Info}(s,u) > 0$; that the *pragmatic content conveyed by utterance u* is the set $\{s \mid \text{Info}(s,u) > 0\}$; and that an uttering of u implicates the negation of $\{s \mid \text{Info}(s,u) < 0\}$. The purpose of sketching these qualitative definitions is merely to underline how the RSA framework makes it possible that a listener computes a pragmatic inference but still puts strong credence in a state that is implicated to be false.

Relation to Gricean maxims. The vanilla RSA model, as described above in Equations (1)–(4), is a direct formalization of Grice's idea that listeners can infer pragmatic meaning based on the assumption that speakers adhere to certain rules of behavior, the so-called "Maxims of Conversation" (Grice 1975). This relationship to Grice's maxims becomes more transparent if we rewrite the speaker's utterance-choice probability, starting from the definition in (3).

(5)

$$\begin{aligned} P_{S_1}(u \mid s) &\propto \exp\Big[\alpha\Big(\log P_{L_0}(s \mid u) - C(u)\Big)\Big] && \text{[definition (3)]} \\ &= \left(P_{L_0}(s \mid u)\ \frac{1}{\log C(u)}\right)^{\alpha} && \text{[rules exponential \& log]} \\ &= \left(\frac{[\![u]\!](s)}{|\,\{s \mid [\![u]\!](s) = 1\}\,|}\ \frac{1}{\log C(u)}\right)^{\alpha} && \text{[definition (1) \& uniform priors]} \end{aligned}$$

The reformulation in (5) shows how the speaker's choice probabilities are a product of three factors, each corresponding to a central postulate concerning how cooperative and, arguably, rational speakers should tailor their contributions to a conversation. Suppose we recast the components as follows: (i) we realize that the expression $[\![u]\!](s)$ just gives us the truth-value of u in state s, so we write this explicitly as $\text{Truth}(u,s) = [\![u]\!](s)$; we identify the expression $|\{s \mid [\![u]\!](s) = 1\}|^{-1}$ as a measure of the logical strength of u, because it gives us the number of states in which u is true, so that we explicitly write this quantity as $\text{Informativity}(u) = |\{s \mid [\![u]\!](s) = 1\}|^{-1}$; (iii) we consider the cost $C(u)$ of utterance u as a measure of efficiency of economy to optimize during production, and so explicitly write $\text{Economy(u)} = \log C(u)$, using the logarithm for mathematical convenience. We can then rewrite the speaker rule as:[3]

$$P_{S_1(u|s)} \propto \text{Truth}(u,s)\ \text{Informativity}(u)^{\alpha}\ \text{Economy}(u)^{\alpha} \tag{6}$$

These three factors directly capture (a formalization of) the Gricean maxims of Quality, Quantity, and Manner. In effect, the RSA speaker assumes that speakers (i) maximize truth (i.e., they utter no falsehoods if they can select at least one true message; Quality), (ii) maximize informativity all else equal (Quantity), and (iii) maximize utterance economy all else equal (Manner). The more rational a speaker (i.e., the higher the value of the scaling parameter α), the more pronounced the latter two optimization tendencies become. This behavior reflects the special status that Grice bestowed on the Maxim of Quality: (binary) truthfulness is not affected by the speaker's degree of optimizing informativity and costs.

Informativity from alignment of beliefs. We saw above that the Vanilla RSA model's definition of speakers' choice probabilities can be interpreted as a product of three factors that correspond closely to Gricean maxims of Quality, Quantity, and Manner. To better understand the motivation behind the speaker production rule in Equation (3), we now show how it can be derived from an intuitive picture of belief alignment. This formal exercise will also immediately pave the way for understanding variations of the RSA model (to be discussed below), which do not assume that the speaker knows the true world state (Goodman & Stuhlmüller 2013, Scontras & Goodman 2017, Herbstritt & Franke 2019).

Suppose that the speaker has some belief about the relevant world states that gets captured in a probability distribution $P_{S_1\text{-}Bel}$; after hearing utterance

3 For binary truth-values, α can be dropped from the factor Truth.

u, the literal listener's beliefs about the relevant world states are represented as another probability distribution $P_{L_0\text{-}Bel}(u)$. An utterance u is more useful than another utterance u' to the extent that $P_{L_0\text{-}Bel}(u)$ (i.e., the listener's beliefs after hearing u) is more closely aligned with $P_{S_1\text{-}Bel}$ (i.e., the speaker's beliefs) than $P_{L_0\text{-}Bel}(u')$ (i.e., the listener's beliefs after hearing u') is. Vanilla RSA corresponds to an implementation of alignment in terms of an information-theoretic notion of divergence between probability distributions. Behind it all is simply the idea that the speaker prefers utterances that best align speakers' and listeners' relevant beliefs about the world.

A useful notion of alignment of probability distributions is the information-theoretic measure of Kullback-Leibler (KL) divergence. KL-divergence is not symmetric, but considers one of two to-be-compared probability distributions as the ground-truth, or the objective function to be approximated. This asymmetry makes sense in language understanding because it is the speaker's beliefs to which the listener should align. KL-divergence then measures divergence in terms of, essentially, the expected number of extra bits needed to encode a signal that was generated from the true distribution with the approximate distribution. The Kullback-Leibler divergence between (baseline/true) probability distribution $P_{S_1\text{-}Bel}$ and (approximate/to-be-optimized) probability distribution $P_{L_0\text{-}Bel}(u)$ is defined as:

$$\mathrm{KL}(P_{S_1\text{-}Bel} \,\|\, P_{L_0\text{-}Bel}(u)) = -\sum_s P_{S_1\text{-}Bel}(s)\ \log \frac{P_{S_1\text{-}Bel}(s)}{P_{L_0\text{-}Bel}(s \mid u)} \tag{7}$$

Using KL-divergence, we can then state a more general definition of utterance utilities, to replace the formulation in (3):

$$U_{S_1}(u;s) = \mathrm{KL}(P_{S_1\text{-}Bel} \,\|\, P_{L_0\text{-}Bel}(u)) - C(u) \tag{8}$$

The formulation in (3) is a special case of the more general (8) that follows if the speaker's beliefs about the relevant world states are degenerate, that is, whenever the speaker knows the true world state s^*, so that $P_{S\text{-}Bel}(s^*) = 1$. In that case, we have:

$$\begin{aligned}\mathrm{KL}(P_{S_1\text{-}Bel} \,\|\, P_{L_0\text{-}Bel}(u)) &= -\sum_s P_{S_1\text{-}Bel}(s)\ \log \frac{P_{S_1\text{-}Bel}(s)}{P_{L_0\text{-}Bel}(s \mid u)} \\ &= -\log \frac{1}{P_{L_0\text{-}Bel}(s^* \mid u)} = \log P_{L_0\text{-}Bel}(s^* \mid u)\end{aligned}$$

This derivation of the speaker production rule in Equation (3) in terms of belief alignment shows how to generalize the Vanilla RSA model to contexts in which the speaker might not be fully knowledgable about the true world state.

3 Variations on Vanilla

We have seen how Vanilla RSA can be used to model pragmatic reasoning in simple reference games, as well as specificity implicatures more generally. However, pragmatic reasoning involves much more complexity than Vanilla RSA captures. In particular, there are many different aspects of uncertainty that must be included in order to handle a broader range of language phenomena. In this section, we explore recent extensions of RSA meant to handle such complex language phenomena; each variation introduces reasoning about different sources of uncertainty:

- contextually-relevant semantic parameters (e.g., degree thresholds; Section 3.1),
- the Question Under Discussion (Section 3.2),
- aspects of the linguistic context (e.g., what counts as relevant or expected; Section 3.3),
- the speaker's access to knowledge (Section 3.4), and
- the speaker's utility calculus (e.g., whether it concerns informativity, politeness, etc.; Section 3.5).

We organize our discussion around deviations from Vanilla RSA, highlighting novel technology in language-understanding computation and the phenomena it captures. While we hope to cover a good deal of ground in this brief overview, it is important to note that the RSA modeling framework is an active area of research with new developments continually arising.

As a point of reference for Vanilla RSA, we repeat the S_1 utterance-choice rule below in (9), where a pragmatic speaker selects utterances in proportion to their utility in conveying some observed state of affairs to a naive listener who interprets the utterance according to its literal semantics—all while minimizing utterance cost. As we shall see, many of the extensions of RSA can be characterized in terms of their deviation from this simple utterance-choice rule.

(9) $$P_{S_1}(u \mid s) \propto \exp(\alpha(\log P_{L_0}(s \mid u) - C(u))), \text{ where}$$
$$P_{L_0}(s \mid u) \propto P_{L_O}(s) \cdot [\![u]\!](s)$$

3.1 Meaning inference

Vanilla RSA assumes that the speaker and listener possess the same model of language, with a fixed, context-independent utterance semantics that is shared by both speakers and listeners. In this way, the speaker and listener are assumed to understand language in the same way regardless of the context in which the language is used. But what happens when there is uncertainty about the meanings of words, either because they are sensitive to context or because speakers and listeners are uncertain about the other's language model, such that the meaning of words might not be shared across the two? We can capture this flavor or uncertainty in RSA by parameterizing the interpretation function such that the utterance semantics is subject to change. This move is illustrated in (10), where the variable **x** modulates the truth-functional mapping for an utterance u. Once named, this interpretation-fixing variable **x** becomes subject to active pragmatic reasoning, resolved by the revised L_1 in (11). In other words, we can allow for uncertainty around **x**, such that the precise semantics of u might be vague or underspecified in context. What results is the class of lexical uncertainty RSA models.

(10) $$P_{S_1}(u \mid s, \mathbf{x}) \propto \exp(\alpha(\log P_{L_0}(s \mid u, \mathbf{x}) - C(u))), \text{ where}$$
$$P_{L_0}(s \mid u, \mathbf{x}) \propto P_{L_O}(s) \cdot [\![u]\!]^{\mathbf{x}}(s)$$

(11) $$P_{L_1}(s, \mathbf{x} \mid u) \propto P_{S_1}(u \mid s, \mathbf{x}) \cdot P(s) \cdot P(\mathbf{x})$$

Lassiter & Goodman (2013) use this technology to model the interpretation of vague gradable adjectives, whose meaning depends crucially on properties of the context, both linguistic and extra-linguistic. Take the adjective *heavy*, which is true of some state (i.e., a weight) just in case the weight exceeds the relevant threshold for heaviness. However, we conclude that a *heavy elephant* weighs substantially more than a *heavy backpack*; moreover, the weight we attribute to the backpack is likely to vary depending on whether the speaker is a preschooler or an olympian. Thus, the threshold above which some state counts as heavy is unlikely to be a fixed, context-invariant value.

Lassiter & Goodman handle the uncertainty around the threshold value by parameterizing the meaning of utterances containing gradable adjectives: *heavy* will be true of some state just in case the weight exceeds the relevant threshold (as before), where this threshold (**x** in (10)) gets fixed by pragmatic reasoning:

(12) $$[\![\textit{heavy}]\!]^{\mathbf{x}} = \lambda s.\textit{weight}(s) > \mathbf{x}$$

It is up to the pragmatic listener, L_1, to resolve the uncertainty surrounding $\mathbf{x}$. To do so, L_1 hears the utterance and jointly infers both the state of the world (i.e., the relevant weight) and the relevant heaviness threshold. L_1 performs this inference just as in the vanilla model: which state and threshold would have been most likely to lead S_1 to select the utterance that L_1 encountered? In other words, L_1 simulates S_1's behavior for the possible combinations of states and thresholds, then weights states and thresholds in proportion to the probability that they would lead S_1 to choose the utterance that was encountered.

A similar threshold-inference mechanism can be used to formalize the truth conditions of generic statements (e.g., *Birds fly*), whose relation to the within-category property prevalence (i.e., the % of birds that fly) has been historically hard to pin down (e.g., not all birds fly). Similar to the gradable adjectives model, the generics model adopts a threshold semantics where the generic is true just in case the prevalence of the feature f within the category k exceeds the relevant threshold:

$$[\![generic]\!]^{\mathbf{x}} = \lambda f.\lambda k.P(f \mid k) > \mathbf{x} \tag{13}$$

Then, pragmatic reasoning can be used to jointly determine the prevalence $P(f \mid k)$ as well as the threshold $\mathbf{x}$ (Tessler & Goodman 2019).[4]

This technology—parameterization of the utterance semantics subject to pragmatic reasoning—allows for a recasting of the division of labor between semantics and pragmatics. With lexical uncertainty, the semantic content of an utterance can be (at least partially) determined via pragmatic inference. Bergen et al. (2016) seize on this innovation to model M-implicatures (Horn 1984) and certain embedded implicatures (Hurford 1974, Chierchia et al. 2012). Rather than uncertainty around the semantics of a single utterance (e.g., a gradable adjective, as in the example above), Bergen et al. assume uncertainty around entire lexica. The lexical uncertainty approach to embedded implicatures is picked up, extended by, and evaluated alongside empirical data by Potts et al. (2016) and Franke & Bergen (2020).

Meaning inference also serves as a useful tool for modeling ambiguity resolution. Scontras & Goodman (2017) use an utterance-choice rule as in (10)

4 Tessler & Goodman (2019) actually use a reduced version of this model where the threshold inference occurs at the level of L_0, as opposed to L_1. Goodman & Lassiter (2015) argued via simulation that threshold inference at the level of L_0 results in interpretations that are too weak. Their simulations, however, did not consider the kinds of state priors that are relevant for generics (see Section 5.1 for further details on state priors for generics). It remains an open empirical question whether threshold inference at the level of the literal vs. pragmatic listener is preferred. For related conceptual criticism of the approach of Lassiter & Goodman (2017), see also Qing & Franke (2014).

to model the resolution of distributive-collective ambiguities in plural predication. If a listener hears that *the boxes are heavy*, the listener must jointly infer both the intended interpretation of the utterance (i.e., distributive, commenting on individual box weight, or collective, commenting on the total weight of the set) and the state of the world (i.e., the weights of the boxes). Returning to (10), Scontras & Goodman treat **x** as the interpretation-resolving variable that determines whether u receives a distributive or collective interpretation. Savinelli et al. (2017, 2018) use this technology to model the resolution of scope ambiguity, as in utterances like *every horse didn't jump over the fence*. For Savinelli et al., **x** in (10) serves to determine the relative scope of logical operators elements at LF. Both of these applications—plural predication and quantifier scope ambiguity—highlight how meaning inference in RSA can serve as a useful shorthand for providing computational-level descriptions of the process of ambiguity resolution. However, while these models offer a promising hypothesis for how listeners reason pragmatically to resolve ambiguity, it bears noting that neither application attempts to model the compositional processes that give rise to the relevant ambiguities in the first place. We return to this issue in Section 6.

3.2 QUD inference

The technology we have considered up to now allows for the pragmatic enrichment of utterances: from "blue" to "blue square" or from vague "expensive" to "expensive" with a contextually-determined price threshold. However, given that S_1's utility relies on successfully communicating the observed state to L_0 (i.e., on informativity), our enrichments can only ever be literal—we have no way for "blue" to be interpreted as "green".[5] But many instances of everyday language use are, strictly speaking, non-literal. For example, if you hear that someone paid "a million dollars" for their coffee at some hipster hangout, you are unlikely to infer that the actual price paid was $1,000,000. Rather, you infer that the price was high and that the speaker is, to put it mildly, less than thrilled.

To allow for enrichments beyond the literal meaning, we must broaden S_1's goals beyond informativity with respect to s. One way of conceiving of

5 Formally, this outcome is a direct consequence of grounding the speaker's production utilities in information theoretic surprisal and KL-divergence, as described in Section 2. The Vanilla RSA model, as based on rules (3), entails that if $[\![u]\!](s) = 0$, while the speaker's belief assigns positive probability to s, there is zero probability of the speaker choosing utterance u. In other words, Vanilla RSA predicts that the speaker never chooses an utterance when the speaker is not certain that the utterance is true of the relevant state.

speaker goals is in terms of the Question Under Discussion (QUD) the speaker aims to answer. Under certain theories of pragmatics, all discourse proceeds with respect to some QUD (Roberts 2012); utterances in the discourse must at least partially answer the QUD in order to be cooperative and felicitous. By allowing for a broader range of QUDs and uncertainty around which QUD is intended, Vanilla RSA may be extended to capture uses of non-literal language.

As was the case with meaning inference in the previous subsection, we can illustrate this innovation by highlighting its effect on the vanilla utterance-choice rule. In (14), S_1's utility continues to break down into informativity and economy components. However, now $\mathbf{x}$ determines what it is that S_1 endeavors to communicate to L_0. Viewing $\mathbf{x}$ as the QUD, $F(s, \mathbf{x})$ serves to map s to the answer to $\mathbf{x}$ that S_1 would like to communicate to L_0.

(14) $P_{S_1}(u \mid s, \mathbf{x}) \propto \exp(\alpha(\log P_{L_0}(F(s, \mathbf{x}) \mid u) - C(u)))$, where
$P_{L_0}(s \mid u) \propto P_{L_O}(s) \cdot [\![u]\!](s)$ and
$F(s, \mathbf{x})$ is a set of states

Kao et al. (2014b) use this technology to model hyperbole, as in the coffee-price example above. To see how, consider some possible states of the world about which a speaker might want to communicate:

(1) *Possible states of the world for the coffee-price scenario*:
s_1: {`affect`: positive, `price`: \$3}
s_2: {`affect`: positive, `price`: \$7}
s_3: {`affect`: positive, `price`: \$1,000,000}
s_4: {`affect`: negative, `price`: \$3}
s_5: {`affect`: negative, `price`: \$7}
s_6: {`affect`: negative, `price`: \$1,000,000}

Notice that the world states in (1) have two properties: the `affect` of the speaker (i.e., whether the speaker feels positive or negative about the price paid), together with the actual `price` paid. A speaker talking about s might have the goal of communicating about their `affect` and the `price` paid, or they may want to communicate only their `affect` or only the `price` paid. Viewed in terms of QUDs, these goals serve to partition the state space. With an `affect?` QUD, we partition the state space into two cells, corresponding to positive vs. negative `affect`. With a `price?` QUD, we partition the state space into three cells, corresponding to the three possible `prices`. With a QUD asking after both `affect-and-price?`, we partition the state space into the six cells listed in (1).

With the utterance-choice rule in (14), S_1's goals are QUD-specific: for communication to succeed, L_0 must simply arrive at the correct cell of the relevant partition; whatever L_0 infers about what happened within that partition is irrelevant to S_1's goals. For example, if S_1 has a negative affect and is addressing the `affect?` QUD, utility depends on whether or not L_0 correctly arrives at the negative cell of the `affect?` partition, regardless of whether L_0 infers the full state to be s_1, s_2, or s_3. It is at this point where non-literal language suddenly becomes rational for a speaker. To see how, we have to say more about the prior knowledge speakers and listeners bring to bear on their communication scenarios.

World knowledge tells us that states in which the speaker actually paid \$1,000,000 (i.e., s_3 and s_6) are extraordinarily improbable, while states in which the coffee cost \$3 (i.e., s_1 and s_4) are the most likely; this knowledge gets reflected in L_1's prior over world states, $P(s)$. We also have prior knowledge about how probable a speaker is to have a positive vs. negative affect given a specific price paid for a cup of coffee: as prices increase, the probability of negative affect increases along with it. Now, return to a speaker addressing the `affect?` QUD: if S_1's primary objective is to communicate negative affect to L_0, and, crucially, if the available utterances only concern price (i.e., "I paid \$3/\$7/\$1,000,000"), then the rational thing for S_1 to say is that the coffee cost "\$1,000,000"; given the extremely strong association between \$1,000,000 coffee and negative affect, L_0 is all but certain to arrive at the correct answer to the QUD: the speaker holds negative affect.

We have identified how non-literal utterances can be useful to a speaker, but we have yet to capture the non-literal aspect of their meaning. It is at the level of L_1 where utterances suddenly become non-literal. Hearing that the speaker paid "\$1,000,000", L_1 uses the informative priors on prices and their associated affects to arrive at a plausible interpretation. Given that coffee is sure to cost less than \$1,000,000, L_1 considers the possible QUDs the speaker might be addressing when selecting u. The `price?` and `price-and-affect?` QUDs are highly unlikely, given that \$1,000,000 is a highly unlikely price to pay for a cup of coffee. That leaves the `affect?` QUD. So, L_1 considers the possible full states that would have led S_1 to utter "\$1,000,000" in response to the `affect?` QUD. Via this counterfactual reasoning, L_1 arrives at a reasonable interpretation: S_1 is likely to have paid more than usual, so s_1 and s_4 are downweighted, but we are still on earth, so s_3 and s_6 remain highly unlikely. Of the remaining states, S_1 is most likely to have wanted to communicate negative affect with "\$1,000,000", so L_1 concludes that s_5 is more probable than s_2. In other words, L_1 hears that the speaker paid "\$1,000,000" for a cup of coffee and concludes (i) that the speaker is likely addressing the `affect?`

QUD, (ii) that the speaker is likely trying to communicate negative affect, and (iii) that the price is higher than usual (which led to the negative affect) but still within the realm of possibility. This last conclusion—a hyperbolic interpretation of the price—is where non-literal meaning enters.

Kao & Goodman (2015) use an extension of the hyperbole model to capture ironic interpretations of weather descriptions (e.g., "It's terrible out" to describe a beautiful sunny day). In addition to inferring speaker affect, the irony model incorporates an additional dimension of meaning that a speaker might want to communicate: their arousal about the weather (i.e., whether they feel strongly about it). The non-literal interpretation mechanism remains the same: if a speaker describes the weather as "terrible" but we know that terrible weather is extremely unlikely (the authors situate their imagined conversation in California, where the weather is rarely objectionable), then a listener will not conclude that the unlikely literal state (i.e., terrible weather) holds, but rather that the speaker must be addressing some other dimension of meaning, namely the strong arousal they feel in their positive affect toward the normal state of affairs: good weather. Similarly in Kao et al. (2014a), who model the interpretation of metaphors like "John is a whale". Rather than inferring that John is an aquatic mammal (a highly unlikely state of affairs), the listener concludes that the speaker chose the utterance to communicate about some other dimension of meaning, for example John's physical stature, which has whale-like properties.

3.3 Context/Prior inference

We now have technology to model vagueness and ambiguity on the basis of an underspecified semantics, as well as shifting goals and the non-literal language that addresses them on the basis of uncertainty around the QUD. So far, all of this reasoning has happened with respect to a fixed context, or common ground. What happens when a listener is unsure of the context a speaker has in mind in the production of their utterance? For example, upon hearing that a person is "tall", how does a listener identify the relevant comparison class against which to evaluate tallness? Does the speaker mean tall for a young child or tall for a basketball player?

In an analogous manner to the way we treat uncertainty regarding the literal meaning and uncertainty regarding the QUD, a listener can have uncertainty about (aspects of) the relevant context. In the case of "tall", perhaps the listener is unsure of the appropriate comparison class. Or maybe the listener is unsure of the facts of the world: what is the general prevalence

of tall-ness? In (15) we amend the speaker choice function to reason about $\mathbf{x}$, a variable that influences the relevant/likely states of the world s.

(15) $$P_{S_1}(u \mid s, \mathbf{x}) \propto \exp(\alpha(\log P_{L_0}(s \mid u, \mathbf{x}) - C(u))), \text{ where}$$
$$P_{L_0}(s \mid u, \mathbf{x}) \propto P_{L_0}(s \mid \mathbf{x}) \cdot [\![u]\!](s)$$

Viewing $\mathbf{x}$ as the comparison class for a gradable adjective, we can build on the Lassiter & Goodman adjectives model mentioned in Section 3.1 to describe the formal mechanism by which listeners reason about the context against which an adjective like *tall* receives a context-specific interpretation. The adjectives model fixes the vague meaning of a gradable adjective by computing the relevant threshold via reference to the state prior, the prior distribution over heights within some comparison class. In a model that incorporates uncertainty about the comparison class, the prior over world states is treated as a conditional probability distribution that depends upon the comparison class $\mathbf{x}$: $P_{L_0}(s \mid \mathbf{x})$.

Tessler et al. (2017) use this parameterization to model comparison class inferences when listeners hear only an adjective applied to an individual (e.g., when describing a basketball player, "He is tall"). The model assumes that the listener has access to the fact that the referent is a basketball player and knows that a basketball player is a person. Then, the listener's prior distribution of comparison classes is the set of all categories of which the referent is a member, weighted by the prior probability. The prior probability of different comparison classes is an unknown theoretical construct and the original comparison class inference model only uses two comparisons classes: the subordinate category (e.g., basketball players) and a basic or superordinate category (e.g., people). The inference proceeds by imagining the conditions under which a speaker would use the adjective heard (e.g., *tall*) with each of the comparison classes (e.g., *for a person* vs. *for a basketball player*).

The uncertainty over the comparison class is a special case of uncertainty over contexts, about which an appropriately-constructed L_1 model can reason. Degen et al. (2015) use this technology to model the stubbornness of scalar inferences in the face of strong prior beliefs that would have made the implicature-calculated meaning false. In that study, the authors investigated listener interpretations of quantified statements involving "some" in a set of stimuli with diverse prior expectations concerning the probability of the *all-state*. For example, "John threw 15 marbles into the pool. Kevin, who observed what happened, said *Some of the marbles sank.*" In this scenario, the prior probability of a marble sinking in water is close to 1; thus, one

would expect *a priori* that all of the marbles would sink. However, the authors observe across multiple dependent measures that participants still draw the implicature—*some but not all of the marbles sank*. Degen et al. account for this behavior using the context-inference mechanism that allows listeners to revise their prior beliefs when the utterances heard are sufficiently unexpected, positing a world that is somehow strange or *wonky*—for example, a world in which marbles sometimes float.[6]

3.4 Epistemic inference

By now it should be clear that a productive way of extending the RSA framework is to take into account various sorts of uncertainty that enter into communication scenarios: uncertainty about the semantics, the goals, or the context itself. Another common source of uncertainty concerns the epistemic states of the agents who are communicating. Speakers commonly describe states of the world without total knowledge of those states. For example, a speaker might observe an empty plate on John's desk with pizza crumbs on it and describe the scenario with "John ate some of the pizza." As far as the speaker knows, John could have eaten all of the pizza, but, without seeing the pizza box, the speaker lacks evidence to make this stronger claim. A hungry listener will interpret the speaker's utterance to infer just how much pizza John has eaten. Importantly, whether the speaker saw the pizza box or just the crumb-covered plate is likely to influence the listener's inference about the full state of the pizza. The mechanism involves reasoning about the knowledge that the speaker used to make her utterance (i.e., the speaker's epistemic state), and the implications that the knowledge has for likely states of the world.

We can model this inference by once again amending our speaker model. Section 2 already introduced a way of extending the speaker production rule in Equation (3) to also cover the case where the speaker is uncertain about the world state. To allow the pragmatic listener to reason in a structured manner about the speaker's beliefs, we model the speaker's belief state $P_S(s \mid \mathbf{x})$ as a function of some observation $\mathbf{x}$ the speaker may have made to arrive at her

6 If we construe the pragmatic content of an utterance of "some" as in Section 2 in terms of the degree to which the utterance changes the pragmatic listener's beliefs about each state, it is still feasible to speak of a "some but not all" implicature of "some" even with heavily biased beliefs. In effect, this means that we should understand the *wonky worlds* model not as an explanation of how the implicature persists in the light of heavy prior biases, but rather as addressing the mechanism by which, on top of the usual implicature, prior beliefs about the world can be revised.

beliefs. The following formulation is equivalent to the previous formulation of speaker utilities in terms of KL-divergence, thus similarly motivated by the idea that the speaker favors utterances based on how closely they help align the beliefs of a literal listener with their own:

$$P_{S_1}(u \mid \mathbf{x}) \propto \exp(\alpha \mathbb{E}_{P(s|\mathbf{x})}(\log P_{L_0}(s \mid u) - C(u))) \tag{16}$$

Goodman & Stuhlmüller (2013) use this approach to model rates of scalar implicature for the quantifier "some." In the pizza example, the relevant inference concerns how likely it is for the listener to strengthen the "some" utterance to "some but not all." Goodman & Stuhlmüller hypothesize that rates of scalar implicature should decrease as the speaker has less information that would verify whether the all-state obtains (e.g., whether John in fact ate all of the pizza); the authors present behavioral data supporting this claim. The amended RSA model offers an articulated analysis of why speaker knowledge access (and listeners' awareness of it) should have the effect it does on interpretation.

In the case of full knowledge access, L_1 reasons upon hearing "some" that if the speaker had wanted to communicate the all-state, she could have said "all;" but the speaker did not utter "all," so she must know that the all-state does not hold. This reasoning leads L_1 to reassign the probability that would have been assigned to the all-state to the other states compatible with the semantics of "some," namely the some-but-not-all states. This decrease in posterior probability assigned to the all-state serves as our index of scalar implicature. With partial knowledge access, the speaker may believe she is in the all-state, but she can never know for sure; similarly, upon observing only a plate with pizza crumbs, the speaker will never know that the all-state (i.e., that all the pizza was eaten) does *not* hold. L_1 knows this about the speaker, and therefore is aware that, when hearing "some" from a speaker with partial knowledge access, the speaker is less likely to be in position to know that the all-state does not hold, thus canceling the counterfactual reasoning mechanism driving scalar implicature.

In their model of distributive-collective ambiguities in plural predication, Scontras & Goodman (2017) use this same sort of speaker knowledge manipulation to modulate rates of distributive interpretations. Hearing "the boxes are heavy" from a speaker who was unable to access the weights of individual boxes—because, for example, the speaker used a dolly to move all the boxes at once, and so accessed only their collective weight—a listener is less likely to interpret the utterance distributively, as commenting on the individual box weights. Rather, the listener is more likely to assign the utterance a collective interpretation, commenting on the total weight of the

set of boxes. The reasoning comes straight from Grice (1975): "Do not say that for which you lack adequate evidence." In Scontras & Goodman's model, the pragmatic listener is reasoning about a speaker who lacks the knowledge to verify a distributive interpretation on the basis of the (partial) observation she has made about the state of the boxes.

Once we have a model of variably uncertain speakers as in Equation (16), it is also feasible to investigate the pragmatic listener's inferences about the likely observation $\mathbf{x}$ that helps explain the speaker's utterance. Conceptually, this task amounts to inferring the speaker's level of expertise or competence. While non-probabilistic models of Gricean inference often make reference to a monolithic (though defeasible) assumption of speaker competence (Geurts 2010, Schulz & van Rooij 2006, Spector 2006), holistic probabilistic inference allows a more gradient picture. For example, after an utterance of a logically-stronger expression like "all", the pragmatic listener will put more credence in the proposition that the speaker has made an observation $\mathbf{x}$ that makes the speaker knowledgable, than they would after hearing a logically-weaker expression like "some".

3.5 Complex utility/utility inference

So far, the different variations of RSA that we have considered all proceed with the assumption that speakers want to convey information to the listener; their utility concerns informativity with respect to the QUD.[7] Though informational uses of language are arguably the bread-and-butter of a successful communicative system, speakers do not always say what they mean. We dawdle, prevaricate, and sometimes outright lie, and this behavior is often in the service of maintaining our social relationships. Perhaps rather than telling your friend that their haircut is ugly, you could say "it's an interesting look". Why do speakers deviate from choosing the maximally-informative utterance? We can model non-informational uses of language by augmenting the speaker's utility function:

$$P_{S_1}(u \mid s, \mathbf{x}) \propto \exp(\alpha(\mathbf{x} \cdot \log P_{L_0}(s \mid u) + (1 - \mathbf{x}) \cdot \mathbb{E}_{P_{L_0}(s|u)}[V(s)] - C(u))) \tag{17}$$

Now, S_1 chooses an utterance on the basis of the true state s (e.g., the speaker's true feelings about the haircut), where the standard RSA information utility is one component of the utility function. This informational

7 Generalizations of the speaker's goals for communication exist. We could, for example, model a speaker's utility function in terms of closeness to the true state, for example, for state spaces with a clear similarity ordering (Franke 2014b).

utility is weighted by a mixture parameter $\mathbf{x}$, where the second utility component receives the opposite weight $(1-\mathbf{x})$. The second component is what Yoon et al. (2016) call *social utility*, and is a function of the literal listener's subjective value V of the state of the world they believe they are in given the utterance. Yoon et al. (2016) deploy this generalization of speaker utility to account for polite language use. It formalizes a version of Brown & Levinson's (1987) politeness theory, in which cooperative speakers have social goals to minimize any potential damage to the hearer's (and the speaker's own) self-image (called *face*), in addition to standard informational goals. Thus, S_1's utility is a mixture of two-utilities: standard epistemic utility defined in (3) and a social utility: $U_{social}(u) = \mathbb{E}_{L_0(s|[\![u]\!])}[V(s)]$. This social utility term increases S_1's production probabilities for utterances that convey states with high subjective value (e.g., "it's *beautiful*", "your talk was *amazing*"), thereby incentivizing white lies.

The pragmatic consequences of this complex utility function can be realized at the higher levels of pragmatic recursion. The pragmatic listener L_1 can reason jointly about the state of the world and about the speaker's utility function (specifically, what is the speaker's trade-off between informativity vs. social goals?). Yoon et al. (2016) show that with information about the true state of the world and the utterance, the model can recover the speaker's likely goals, and, vice versa, with information about the speaker's goals and the utterance, the model recovers the likely true state of the world.

4 Pragmatic agents as probabilistic programs

Our exposition of the framework so far has centered on descriptions of the models in terms of mathematical notation, more concretely in terms of formulae characterizing conditional probability functions (mass or density, depending on the case). We tried to motivate and explain the ideas behind these formulae, for example by showing how the speaker's utility function in the Vanilla RSA model can be derived from a picture of communication in which the speaker is trying to minimize the distance (measured in information-theoretic terms) between the speaker's and a literal listener's beliefs.

While it is most common to communicate a probabilistic model in terms of such formulae-based representations in scientific writing, this approach is not the only way of conceptualizing a probabilistic model (be it an agent model, a statistical model, or otherwise). Indeed, one very intuitive way of thinking about a probability distribution focuses not on a high-level, abstract description of the overall probability of events, but on the *stochastic*

generative process that we conceive of as the source of these events. The web-book "Probabilistic Language Understanding", for which this paper is an introductory and complementary companion, uses probabilistic programs to implement RSA-style probabilistic agent behavior in terms of explicitly-formalized stochastic generative processes (Scontras et al. electronic). The main motivation of this section is to: (i) highlight the distinction between "thinking in math" and "thinking in code" when it comes to developing probabilistic models, (ii) to demonstrate the usefulness of "thinking in code" and thereby underlining the added value of the web-book, and (iii) to argue that, ideally, researchers should develop models by flexibly exploring conceptually-reasonable ideas in terms of both a mathematical and a code-based algorithmic perspective.

To see the difference between a mathematical and a process-based representation of a probability distribution, consider the following definition of a probability density function on some $x \in [a;b]$ parameterized by some integer-valued $n > 0$:

$$P(x\ ;\ n) = \begin{cases} \sum_{k=0}^{n} (-1)^k \binom{k}{n} \left(\frac{x-a}{b-a} - \frac{k}{n}\right)^{n-1} \operatorname{sgn}\left(\frac{x-a}{b-a} - \frac{k}{n}\right) & \text{if } x \in [a;b] \\ 0 & \text{otherwise.} \end{cases}$$

Unless you are already intimately familiar with the Bates distribution, chances are that this mathematical representation tells you very little about what is captured here. Yet, despite the complex math, the Bates distribution derives from a very simple procedural idea: sample n times from a uniform distribution in the interval $[a;b]$ and look at the mean x over all of these n values. The Bates distribution gives the probability of sampling any concrete value of x by that process. The upshot of this example is that a formula-based mathematical representation of a probability distribution can be highly useful for some applications, but it might not be useful for all purposes, one such purpose being to think about a possible process that could generate a particular output in a certain (stochastic) manner.

Probabilistic programming languages like WebPPL (Goodman & Stuhlmüller 2014), in which the examples from the accompanying web-book are written, help to make the process-based perspective explicit. Complex probability distributions, even those for which no closed-form mathematical representation is known, can be approximately represented by a large-enough set of representative samples. These approximate representations are good enough for most common practical purposes. Even more useful than a large set of fixed samples is a function that returns samples every time it is called. Considerations like these are why probabilistic programming languages

like WebPPL allow the user to define probability distributions in terms of a *sampling function*, formulated simply as a process that would generate a single sample (if executed once). For example, the Bates distribution could be written down in WebPPL-inspired pseudo-code as follows:

```
var Bates_sampling_function = function(n, a, b) {
  // take n samples from a uniform distribution on [a;b]
  var unif_samples = sample_from_uniform(n, a, b)
  return(mean(unif_samples))
}
```

This sampling function is sufficient to work with the Bates distribution (at least for most common applications). Indeed, the web-book defines the RSA speaker and listener protocols, which we explicated above in terms of formulae, in exactly these kinds of sampling functions.

Adopting a perspective on language production and interpretation in terms of sampling functions is what we here refer to as "thinking in code." To "think in code" is particularly useful in the context of cognitive modeling because it enables a sampling-based perspective of an agent's reasoning and decision-making process in line with recent accounts of resource-bounded rationality (Griffiths et al. 2015, Lieder & Griffiths 2020). We will walk through two example cases where "thinking in code" and the adoption of a resource-bounded rationality perspective can add interesting flexibility to the modeling of pragmatic agents.

4.1 Utterance priors & utterance salience

The Vanilla RSA model defines the speaker's utterance-choice probability in terms of *utterance costs*, $C(u)$, which we identified as a potential formalization of Gricean Manner or a general constraint of *utterance economy*. Here is the original definition and the suggested refactorization from Section 2:

$$\begin{aligned} P_{S_1}(u \mid s) &\propto \exp(\alpha(\log P_{L_0}(s \mid u) - C(u))) \\ &= \text{Truth}(u,s)\ \ \text{Informativity}(u)^{\alpha}\ \ \text{Economy}(u)^{\alpha} \end{aligned} \tag{18}$$

In this formulation, the cost $C(u)$ of an utterance u affects the speaker's choice of expression: the higher the cost of u, all else equal, the lower the probability that the speaker picks this utterance. This design is motivated by the idea that speakers economize production effort, preferring to utter shorter or otherwise more efficient words or phrases. Consequently, in line with this interpretation of minimization of utterance effort, the strength of the impact of utterance costs is modulated by α: the higher the value of α, the

stronger the speaker's tendency to minimize utterance costs, all else equal. This outcome is intuitive if we think of α as a rationality parameter: the more rational an agent, the more sensitive that agent should be about the minimization of effort.

Adopting a sampling-based perspective inspired by resource-bounded rationality, we can conceive of a natural alternative to the speaker model in (18), in which differences between utterances are not treated as costs to be rationally economized, but rather as differences in salience—that is, the ease with which utterances comes to mind. The latter factor of *differential salience* is arguably not subject to optimization: it is not the case that the more rational a speaker is, the more she would tend to select only the utterances that come to mind most easily. We can implement differential salience (ease of retrieval) with a mathematical formulation as follows:

$$P_{S_1}^{\text{Salience}}(u \mid s) \propto \text{Truth}(u,s)\ \text{Informativity}(u)^{\alpha}\ \text{Salience}(u) \tag{19}$$

The speaker model in (19) includes the factor Salience(u) as an *utterance prior*, suggesting an algorithmic picture of utterance choice: the speaker searches for utterances to choose, based on a gradient of how easily these utterances come to mind, then compares the available options (weighted by their salience/prominence) based on the other factors relevant for communication: truth and informativity (see van Tiel et al. 2021 for an example application of salience priors to the choice of quantity expressions).

The salience-based speaker model in (19) is implemented very naturally in WebPPL (or other probabilistic programming languages); WebPPL-inspired pseudo-code looks as follows:

```
var speaker = function(state) {
  // sample an utterance from an utterance prior
  //   (= 'see what comes to mind')
  var utt = sample_from_utterance_prior()
  // weigh sampled utterance by how true and informative it is
  adjust_score(truth(utt, state) * informativity(utt)^alpha)
  // return utterance
  return(utt)
}
```

Notice that, for technical reasons, WebPPL indeed *requires* the specification of an utterance prior even if we do not use it to implement salience.[8]

8 WebPPL requires the inclusion of an utterance prior because it minimally needs to know the set of all available utterances. If no *differential salience* is to be modeled, the utterance prior in a WebPPL program must still be specified but should then just be uniform.

4.2 Uncertainty from sampling

Another example of a case where "thinking in code" makes a plausible conceptual idea much more readily available than a formula-based approach arises when we model a speaker who is uncertain about the true state of the world. In Section 3.4, we described the model of Goodman & Stuhlmüller (2013), the speaker component of which is repeated here:

$$P_{S_1}(u \mid \mathbf{x}) \propto \exp(\alpha \mathbb{E}_{P(s|\mathbf{x})}(\log P_{L_0}(s \mid u) - C(u))) \tag{20}$$

The speaker is assumed to choose utterances proportionally to how much they minimize the Kullback-Leibler divergence between the speaker's own belief and the literal listener's belief after each utterance. This model is easily written out in concise mathematical notation like in Equation 20, but the WebPPL code to implement it is rather clumsy (see accompanying web-book) and its calculation can be very costly for large sets of possible world states. However, "thinking in code" and conceptualizing a resource-limited reasoner suggests a different model, one that is difficult to write down in mathematical notation but that is much easier to write down in a probabilistic program—and also much more efficient to approximate. This alternative speaker model looks as follows in WebPPL-like pseudo-code, where `belief_state` is a probability distribution capturing the speaker's belief state:

```
var speaker = function(belief_state) {
  // sample an utterance from an utterance prior
  var utt = sample_from_utterance_prior()
  // sample a possible world state from the speaker's beliefs
  var state = belief_state()
  // weigh sampled utterance by how true and informative it is
  //   based on the current sample of a possible world state
  adjust_score(truth(utt, state) * informativity(utt)^alpha)
  // return utterance
  return(utt)
}
```

This algorithmic speaker model imagines that, instead of globally reasoning about all possible world states and weighing each one in terms of how probable they are to the speaker, the speaker just samples a random world state from her beliefs and then chooses a true and informative utterance for the sampled state.[9]

9 In fact, it is this easier model that Goodman & Stuhlmüller (2013) implemented, as is apparent from the plots of the model's predictions given in the paper. The key difference between the KL-based and the sampling-based model is that the former will never have an agent choose an utterance for which she cannot rule out that it might be false. See the accompanying web-book for more on this topic.

This example serves to highlight how the available tools—being able to both "think in math" and to "think in code"—help to shape the resulting theories. By having a multiplicity of tools handy, the modeler enlarges the space of conceivable (and practically-realizable) theories. While our focus in the current paper is RSA from a mathematical perspective, much of the work in implementing an RSA model involves conceiving of the relevant agents as sampling-based probabilistic programs; the accompanying web-book offers a hands-on introduction to this skill. In the following section, we dive more deeply into the practical considerations faced in the design and implementation of an RSA model.

5 Modeling practicalities

Section 3 presented a menu of options to the would-be modeler: given a language-understanding phenomenon, which pieces of technology are likely to prove most useful in a formal treatment? Non-literal phenomena might benefit from a treatment via QUD inference or complex utility; vague or underspecified language suggests meaning inference, combined perhaps with uncertainty about the relevant context/prior; in cases where the speaker is unknowledgeable or behaves in an unexpected manner, we can try epistemic inference and/or context inference. Having introduced the basics of the RSA modeling framework, as well as useful additions and amendments, our focus now turns to the practical considerations involved in designing and implementing an RSA model, and testing the predictions of that model against human behavior.

We focus in this section on practical issues arising during model development and testing. Concretely, we will address questions like: which discrete choice options (e.g., world states and utterances) to include in a model; which priors to assign to world states; how to set other (numerical) model parameters (e.g., the rationality parameter α); how to link model predictions and empirical data for (Bayesian) data analysis.

5.1 Determining world states, utterances etc.

One of the most important modeling decisions concerns setting up the conversational scenario: how do we represent the states of the world and utterances that can be described in our model? In other words, what constitutes a world-state, and at what level of linguistic description will utterances be treated? Answering these questions often entails specifying the universe of possibilities of states and utterances (i.e., the support of the prior distri-

butions). In the reference game from Frank & Goodman (2012) depicted in Figure 1, we equate states with individual objects, and the set of possible states corresponds to the set of objects present in the visual array. But only certain features of the objects are relevant: their shape and color. Properties like size or position do not enter into the reasoning about shape and color in this reference game (or at least we do not think they do), and so they need not enter into the representation of states in the model. In (1), we saw possible states from the hyperbole model of Kao et al. (2014b); of all the various facts of the world, the authors narrowed in on two for their model: the price paid for some object and the valence of the affect of the person who paid it. But the authors did not consider all possible prices ranging from \$0 – \$5 trillion (roughly the total amount of money in the world); they narrowed the possible prices to a set that is probably reasonable for the conversational scenario at hand.

One of the more ingenious—and abstract—state implementations comes from the generics model (Tessler & Goodman 2019). Generic statements update beliefs about the prevalence of a feature in a category (e.g., the prevalence of egg-laying among robins upon hearing that "robins lay eggs"). The prior, then, is a distribution over possible prevalences. Tessler & Goodman treat this distribution as a distribution over alternative kinds (e.g., bluejays, horses, tarantulas, etc.), akin to the implementation of the comparison class in the vague adjectives models (Lassiter & Goodman 2013, Tessler et al. 2017). Each kind is then modeled as having its prevalence for the relevant property.

The choice of what possibilities to include is relevant to every aspect of the model where alternatives must be specified: alternative states, alternative utterances, alternative QUDs, etc. There are several options for determining these internal parameters. First, if a model's foremost purpose is the analysis of data from a concrete experiment, then it may be that the stimuli and available choice options of the experiment suggest, if not necessitate, a precise set of states, utterances, or other model ingredients. The experiments of Frank & Goodman (2012), for example, present exactly three objects as potential referents and offer only a small set of discrete choice options for utterances in each trial. So here, and in many other cases, the logical design of the experiment to be modeled dictates the nature and scope of some or all internal parameters.

Without obvious constraints from an experimental setup (or in order to determine those constraints), another reasonable approach is to incorporate an experiment in the model-development pipeline to determine empirically what reasonable state distinctions might be. For example, in developing their model of metaphors like "John is a whale", Kao et al. (2014a) began

by eliciting the most salient features for a set of animals (e.g., *large*, *graceful*, and *majestic* in the case of whales). Another possibility is, when dealing with numerical world states (e.g., in the interpretation of adjectives or other vague expressions), to use a pre-test to elicit likely ranges of numbers for particular experimental items (e.g., Schöller & Franke 2017).

Determining the set of alternative utterances is an open issue in pragmatic theories broadly (Chemla & Singh 2014, Katzir 2007, Fox & Katzir 2011), and the literature on RSA has not yet arrived at a set of best practices for determining alternative utterances. A few general considerations can still be mentioned. First, where linguistic theory makes clear indications of which alternatives are reasonable (e.g., from lexical association or from focus mechanisms), the RSA modeler should take this insight into account. In some domains of application, however, no such clear predictions exist (e.g., for situated referential language, generics, or tropes). In these cases, it is important to minimally ensure that the presented RSA model is *prima facie* plausible as a representation of the speaker's or listener's subjective conceptualization of the decision situation (production or interpretation, depending on what we focus on; Franke 2014a). For example, this perspective on what an RSA model represents can make it plausible to also include an informationless "silent" or "null" utterance. Conceptually, a "null" utterance can be plausible in cases where speakers may reasonably be assumed not to want to say anything at all. Technically, a "null" utterance can be useful for understanding the meaning of an utterance even without reasoning about explicit alternatives (e.g., in the study of vagueness; Lassiter & Goodman 2013).

5.1.1 Prior probabilities of world states

Having decided what the relevant alternatives are and how to model them, the task then turns to implementing the appropriate priors: beliefs about which alternatives are more or less likely a priori. Some priors can be estimated empirically. With state priors, it is sometimes possible to ask people explicitly about the relevant probabilities. For example, Kao et al. (2014b) asked participants to rate the probability of paying various prices for certain objects and how likely it is that someone would consider a given price expensive; those ratings were normalized and served as the state prior probabilities in the hyperbole model. In the reference game model, Frank & Goodman (2012) got at the state prior probabilities using a less direct method. To determine which objects in Figure 1 were more likely to get referenced a priori, the authors told participants that someone had used an

unknown word to signal one of the objects; based on participants' guesses at which object the speaker was intending, Frank & Goodman were able to estimate the relevant state prior. In the experimental tests of the generics model, Tessler & Goodman (2019) employ two alternative techniques: one has participants freely generate categories and then rate the prevalence of a target feature for those categories; another has participants answer targeted questions about the abstract parameters of the prior distribution and then uses Bayesian data analysis to reconstruct the appropriate priors.

In some cases—for example, an analytic demonstration of a qualitative prediction—prior estimation might be unnecessary or infeasible. In such cases, one ought to make minimal assumptions, for example by assuming an uninformative, uniform prior. Better yet, the modeler may want to demonstrate the qualitative predictions under a number of different assumptions about the shape of the prior to show the model's (in)sensitivities to choice of prior.

5.2 Further (numerical) parameters

Like the specific probabilities assigned in the various priors, other free parameters of the model must be fixed in order to generate predictions. Free parameters common to RSA include α, which controls speaker optimality, and the various utterance cost parameters. These parameters are often not of direct theoretical interest but sometimes can be informed by domain knowledge. For instance, it is often natural to assume that the cost of an utterance is a function of either its length (which could be measured in terms of number of words, syllables, or characters) or that the salience of an utterance is related to its frequency (estimated from a corpus).

Unfortunately, no comparable motivation exists for the setting of the speaker optimality parameter α, and efforts to convince a participant that a speaker is more or less rational may lead to unintended pragmatic inferences. For these reasons, the value of α is generally not directly theoretically interpretable. In practice, however, α tends to be greater than or equal to 1, and not very high ($\lessapprox 50$). The interpretation of an α less than 0 is that (the listener believes that) the speaker behaves irrationally, preferably chosing the options which minimize expected utility. If such a value of α is needed to obtain otherwise reasonable model predictions, this is a signal that the model is in some way mis-specified, which may prompt a closer look at the state and utterance parameterization. See Zaslavsky et al. (2020) for a fuller discussion of the role of α in RSA.

For the Vanilla RSA model, the number of layers of recursion can also be thought of as a free parameter. Empirically, it has been observed that levels of recursion and speaker optimality trade-off with one another and hence lead to unidentifiability of the parameters (Frank et al. 2016). Recent analytic analyses have shown, however, that speaker optimality determines a tradeoff between the expected utility optimized by RSA recursion and a conception of communicative effort (Zaslavsky et al. 2020).

5.3 Linking functions

A fully-specified RSA model makes qualitative, probabilistic predictions that can be tested against empirical data. Interpreting those predictions, however, requires a clear linking function from the model output to human behavior, and the specific linking function needed will depend on the human behavior to be modeled.

RSA returns predictions in terms of the probability of either a state and/or other variables inferred by the listener (e.g., QUDs, thresholds, etc.) or a choice of expression or utterance (for speaker models). Thus, if the behavioral data is in the form of a particular state or a particular utterance (e.g., in a forced-choice task), then the model can be related to the data directly; the model provides the probability of the observed outcomes and it is possible to compute statistics such as the likelihood of the data under the model (and compare to the likelihood of the data under different models).

In cases where the response variable is not a particular state or utterance, a linking function will need to be specified. One common response variable for experimental semantics/pragmatics researchers is a truth-value judgment (Crain & McKee 1985, Crain & Thornton 1998). These judgments are used to identify the situations that a sentence can truthfully describe. Assuming that the truth conditions of a sentence—what it takes to make the sentence true—are at least constitutive of sentence meaning (e.g., Chierchia & McConnell-Ginet 2000), then truth-value judgments inform how it is that speakers understand sentences. But how can we model such judgments? Given that truth-value judgments are used to inform our understanding of sentences, there is a temptation to treat the process that goes into generating these judgments as one of language comprehension, corresponding to the listener layer of an RSA model (e.g., Potts et al. 2016). However, several authors have suggested that truth-value judgments are better modeled as a form of language production: whether a speaker would use the sentence to describe a given scenario (e.g., Degen & Goodman 2014, Franke 2016, Savinelli et al. 2017, Tessler & Goodman 2019, Jasbi et al. 2019). In RSA terms, this decision

links to the probability of uttering the sentence of interest given some observed state—predictions from a speaker layer of the model. In models where the S_1 predictions depend on more than merely the observed state (as in the extensions discussed Section 3), generating speaker predictions that correspond to truth-value judgments may require an additional layer of recursion: an S_2 who observes a state s and chooses an utterance to communicate s to L_1, who in turn does the work of resolving any additional variables that S_1's calculation uses (e.g., the interpretation-resolving $\mathbf{x}$ in (10) and (11); Savinelli et al. 2017).

Other common experimental measures include slider-bar or likert-scale ratings (e.g., of the probability of a state) or betting procedures (e.g., eliciting the probability of all states simultaneously as in Frank & Goodman 2012). Though there is the temptation to interpret these measures as directly relatable to the model output (e.g., they may both be treated as probabilities which sum to 1), the correct way of handling these data is by articulating a linking function. For example, likert-scale ratings should ideally be analyzed with a cumulative logit or cumulative probit linking function used for ordinal regression (Franke 2014b, 2016). Or, when a participant places ratings via sliders on all states, a logit-transformed Gaussian may be an appropriate link so that the model places a probability on the actual responses that participants provide (see Franke et al. 2016 for an analysis of different linking functions in the context of a prior-elicitation task).

5.4 Parameter estimation and model comparison

Besides actually building the RSA model of a linguistic phenomenon, the modeler is often tasked with evaluating (or criticizing) the model, as well as comparing it to alternative models. These evaluations and comparisons may be informal when examining whether relevant qualitative phenomena (e.g., symmetry-breaking) are exhibited by a model. Comparisons and evaluations, however, like the modeling of the pragmatic phenomenon itself, can be formalized. In this regards, Bayesian data-analytic methods are regarded as the gold-standard, in addition to being theoretically-synergistic with the Bayesian pragmatic views of language understanding. Instead of a listener asking "what did a speaker likely intend given their utterance (and my internal model of the speaker)?", the question of parameter estimation for the scientist is "what are credible values of the model's parameters given the data we've observed (and our model of the data [i.e., the RSA model])?" Many resources are available for learning Bayesian data analysis; we recommend:

Gelman et al. (2013), Kruschke (2014), Lee & Wagenmakers (2014), Lambert (2018).

For each of the above modeling practicalities, the modeler will find herself with additional decisions about variables of the model. For example, the speaker optimality parameter α is generally considered a free parameter of the model that may be fit to the empirical data (i.e., it adds a degree-of-freedom to the model). The most straightforward Bayesian approach to addressing what value to set for α involves setting a prior over plausible values of that parameter and then performing Bayesian inference, conditioning on the observed behavioral data in order to infer what the most believable values of α are given the RSA model and the observed data. Plausible *a priori* values of parameters can be gleaned by reviewing other related RSA models. For example, at the time of this writing, we observe that most RSA models use α values between 0 and 20, though the actual plausible range of values will differ depending on model architecture (see Tessler & Goodman 2019, Yoon et al. 2020 for some examples). Parameter estimation is particularly interesting when targeting model parameters that are themselves of some theoretical relevance. For example, Yoon et al. (2016) use empirical data to infer plausible values of the parameter $\mathbf{x}$ from Equation (17), which models the relative importance of a speaker's informational and social goals during language production. Schöller & Franke (2017) infer plausible semantic threshold variables for the vague quantifiers *many* and *few*, one pair for each of several different contexts, in order to address the question of whether it is plausible that the same threshold parameter explains the use of these expressions across different contexts of use. Schuster & Degen (2020) perform a similar Bayesian data analysis to infer semantic variables over uncertainty expressions (e.g., *probably*, *might*), finding that listeners entertain quite a bit of uncertainty about the semantic threshold for uncertainty expressions. Degen et al. (2020) infer semantic "noise" values for color and size modifiers, tracing speakers' propensity to over-modify with color adjectives back to differential semantic noise parameters for size vs. color adjectives.

In addition to estimating model parameters, Bayesian methods can be used to arbitrate between competing hypotheses (so-called "model comparison"). At a high level, the approach is the same: we are asking which of two (or, n) models is the most likely explanation of the observed data. In practice, this aim is accomplished using what are called *Bayes Factors* (or BF for short), which compare the average (marginal) likelihood of the observed data under each model (Jeffreys 1961, Kass & Raftery 1995). Here, a "model" is considered both the RSA model and the prior distribution over parameter values (e.g., that α is sampled from a Uniform prior between 0 and 20). Since BFs take into

account the prior distribution over parameters, models with more parameters will have broader prior distributions, and thus the average likelihood of the observed data under such models will intuitively be watered down by these broad priors (e.g., when the model makes very good predictions but only for a small set of values of the parameters). In this way, BFs take into account model complexity (i.e., more complex models will be penalized if the gains in model fit do not compensate for the increased complexity of the model) when deciding which is the best model. Examples of comparisons of RSA models based on empirical data include the work of Potts et al. (2016) (albeit in this case the authors do not use Bayesian model comparison), Qing & Franke (2015), Franke & Bergen (2020), Yoon et al. (2020), and Bohn et al. (2019b).

6 Extensions/limitations

With a better sense of the many aspects of probabilistic language understanding that the RSA framework has been used to capture, we turn now to current limitations of the framework and opportunities for further extensions.

6.1 Compositionality

All of the models we have considered operate at the utterance level, taking as their starting point whatever the compositional semantics delivers to them as the meaning of a proposition; the models deliberately avoid the composition of the literal interpretations over which they operate. This move is made primarily for convenience. However, one of the most remarkable aspects of natural language is its compositionality: speakers generate arbitrarily-complex meanings by stitching together their smaller, meaning-bearing parts. The compositional nature of language has served as the bedrock of semantic (indeed, linguistic) theory since its modern inception; Montague (1973) builds this principle into the bones of his semantics, demonstrating with his fragment how meaning gets constructed from a lexicon and some rules of composition. Since then, compositionality has continued to guide semantic inquiry: what are the meanings of the parts, and what is the nature of the mechanism that composes them? Put differently, what are the representations of the language we use, and what is the nature of the computational system that manipulates them?

We have seen how formalizing pragmatic reasoning in the form of computational cognitive models both informs pragmatic phenomena and enriches theories of semantics. Still, by operating primarily at the level of propositions,

the RSA approach necessarily eschews much of the compositional machinery that generates those propositions in the first place. In principle, this semantic leveling is unnecessary; our models of meaning can take into account the rich compositionality of the communicative system they are meant to characterize. The many sources of uncertainty in semantic composition are ripe for a probabilistic treatment, and we now have the tools to deliver one. The basic functionality we are after is already built into the system: lambda abstraction, functional application, declarative memory and thus the possibility for a lexicon.

Indeed, some initial steps have already been taken toward incorporating compositionality within RSA. Goodman & Lassiter (2015) show how to use the tools of probabilistic programming—the backbone of many RSA applications—to deliver a stochastic lambda calculus that composes the meanings of utterances out of their component parts. Goodman & Stuhlmüller (2014) go one step further, developing a semantic parsing system based on combinatory categorial grammar that exists internal to the L_0 layer of RSA; the parser constructs literal interpretations and verifying worlds from the semantic atoms of sentences. However, neither of these approaches has been scaled up to the full RSA reasoning chain.

By incorporating semantic composition into the RSA framework (rather than approximating it), we further shrink the theoretical and practical distance between semantics and pragmatics by incorporating both within a single model of meaning in language. Explicitly modeling semantic composition may also help us address the largely ad hoc nature of specifying utterance alternatives inherent to the design of current models. By incorporating a generative grammar, the alternative utterances are those that the grammar may generate; and by dynamically constructing the verifying worlds along the way, we may be able to move past hand-coded priors and toward a more general theory of alternatives and state priors. To that end, future work in RSA should examine the ways that a semantic compositional mechanism may be modeled dynamically and probabilistically, within the broader framework of computational cognitive science. A comprehensive approach to modeling language meaning, which treats semantics and pragmatics jointly as a process of probabilistic inference, not only increases the validity of our semantic and pragmatic theories, but more directly informs the psychological underpinnings of language.

6.2 Adaptation

In a Bayesian model, the computation of a posterior distribution can be viewed as a specification of the prior distribution in anticipation of the next update. In other words, "today's posterior is tomorrow's prior". The fact that RSA is a Bayesian model opens the door to formal treatments of adaptation and learning.

A form of adaptation is at play in the formation of conventions. Speakers and listeners must coordinate on how to refer to things in the world. Such convention formation is studied using so-called "tangram" experiments, where participant dyads take turns referring to abstract shapes (the "tangrams") with the intention of getting their partner to pick out the same referent. Over repeated rounds of this game, participants converge on conventional names for the referents, which are often reduced forms from how they originally refered to them (Clark & Wilkes-Gibbs 1986, Hawkins et al. 2020). This lexical adaptation process can be formalized in the RSA framework through interlocutors' reasoning about each others' lexica, as in the meaning-inference models surveyed in Section 3.1 above (Hawkins et al. 2017). Beyond the lexicon, adaptation can occur at the level of pragmatics: listeners may adapt to how others are using language (e.g., what utterances certain speakers prefer). Schuster & Degen (2020) examined listeners' interpretations of uncertainty expressions (*might*, *probably*) and found that listeners update their expectations about the usage of such expressions after minimal evidence from a speaker. Using a model-comparison approach within the RSA framework, the authors found that this kind of semantic/pragmatic adaptation is explained by listeners adapting their representations of the speaker's lexicon (i.e., word–meaning mappings) and the utterances a speaker prefers to say.

6.3 Language development

The RSA modeling framework puts forth a particular hypothesis about how language understanding unfolds, and the model components and parameters can be taken to correspond to a hypothesis about the representations that may vary across the population of speakers or develop with age. For example, at a young age, children are sensitive to various information sources that help them learn the meanings of new words (e.g., an understanding of common ground, informativity, and their extant lexical knowledge; Markman & Wachtel 1988, Schulze et al. 2013, Akhtar et al. 1996). In naturalistic contexts, multiple information sources may be present, raising the question of how these sources are integrated. RSA provides a hypothesis about this integration:

informativity is captured by the speaker rationality parameter; common ground influences the prior over states. Building on a paradigm pioneered by Frank & Goodman (2014), Bohn et al. (2019a,b) interrogated this hypothesis by studying 2-to-5 year old children in word learning tasks. They use the RSA architecture to formalize a space of hypotheses about the integration process and its development (e.g., a rational integration process, or a biased integration process that deviates from RSA's Bayesian logic). They find the developmental trajectories in word learning inferences trace back to the improved reliability of the various information sources and not the integration mechanism itself. Similarly, Savinelli et al. (2017, 2018) use RSA to model child and adult behavior on ambiguity resolution, thereby providing computational evidence in favor of developmental continuity.

Probabilistic speaker and listener behavior, as captured by RSA models, also helps build more realistic scenarios for language change or language evolution. Abstractly put, language learners mainly observe pragmatic language use, but need to reconstruct the underlying semantic structure that likely generated the speech they observed. When this learning or inference task is carried out repeatedly across (stylized) generations, transmission inefficiencies (e.g., through learning biases) can accumulate and shape the cultural evolution of language (e.g. Brochhagen et al. 2018, Carcassi et al. 2019, Carcassi 2020). Woensdregt et al. (2020) apply the combination of iterated learning and RSA modeling to address questions after the potential co-evolution of language and mindreading through numerical simulations. Ohmer et al. (2020) show how pragmatic language use, modelled in terms of RSA speaker and listener protocols, shapes and speeds up the emergence of conventional semantic meaning through reinforcement learning. Lund et al. (2019) use RSA to model the engine of pragmatic reasoning in the diachronic development of linguistic meaning.

6.4 Levels of analysis

In our characterization of RSA reasoning, we have described agents performing various reasoning steps, talking about L_1 reasoning about S_1's reasoning about L_0, etc. However, this language can sometimes prove misleading. RSA models are intended to deliver a computational-level description of the problem speakers face as they use language to communicate. As a computational-level analysis (Marr 1982), these models intentionally avoid pronouncements about the specific mechanisms involved in solving the problem of language understanding. In other words, the RSA framework is not conceived as a model of language processing.

Indeed, one may wonder about the psychological validity of the rational agents operating within an RSA model. Language users face limits on their cognitive abilities—limitations of memory, attention, and general processing resources—yet it may not be obvious where such resource limitations enter into RSA. One avenue through which resource limitations may enter is the optimality parameter α, which control's S_1's utility-maximizing behavior. For example, one may posit that α actually varies across demographic features of the data set (e.g, age of participant); then, one may model α via linear regression (predicting α as a function of age) to see how pragmatic competence develops (Bohn et al. 2019b). As α decreases, S_1 is less likely to choose actions (i.e., utterances) with higher utility. Another avenue through which resource limitations may enter RSA is the number of layers of recursive reasoning (Franke & Degen 2016). Moreover, as detailed in Section 4, by taking direct inspiration from probabilistic programs, we may see sampling approximations to inference processes as another factor in which RSA-like behavior is a form of resource-bounded rationality (Griffiths et al. 2015).

As a computational-level model operating over full utterances, RSA, at least in the formulations we have considered, also eschews the highly incremental nature of human language processing. Rather than waiting for utterances to terminate before reasoning about them, language users begin integrating information as soon as it arrives during the linear unfolding of time (e.g., Tanenhaus et al. 1995). This incremental processing is not limited to word recognition or syntactic parsing; pragmatic reasoning has also been shown to be incrementally engaged (e.g., Sedivy et al. 1999, Sedivy 2007). We might wonder, then, whether RSA should incorporate incremental pragmatic reasoning, and, if so, how such reasoning may be implemented. Cohn-Gordon et al. (2019) offer one path forward for incremental RSA, amending the speaker and listener models so that they reason about individual words and potential continuations, showing how this advance is beneficial for natural language processing applications. In psycholinguistics, Werning & Cosentino (2017) and Augurzky et al. (2019) show how probabilistic predictions about the whole utterance derived from RSA-style modeling can be broken down to next-word expectation, a finding which is supported even by quantitative aspects of EEG measurements.

6.5 Natural language processing

The final extension we discuss concerns the utility of RSA reasoning to current applications in language technologies. Probabilistic formalizations of Gricean pragmatics lend themselves directly to integration into probabilistic

models used in natural language processing. One line of research focuses on how a semantic meaning representation can be learned from data, based on the assumption that the data was generated by an RSA-style pragmatic agent. For example, Monroe & Potts (2015) explore how to set up a model in which RSA-style literal and pragmatic language agents learn underlying semantic meaning associations based on data from the TUNA corpus (van Deemter et al. 2006), which provides human data from tasks similar to the reference game from Section 2. Monroe & Potts show that pragmatic agents outperform literal language users in their set-up. Monroe et al. (2017) apply a similar strategy of learning semantic representations through the lens of a pragmatic agent to a novel corpus of color descriptions.

Another line of recent research that brings together probabilistic pragmatic reasoning and computational linguistics centers around the realization that we can increase system performance by taking into account a pragmatic perspective for generation and interpretation. Early work in this vein comes from Andreas & Klein (2016), who show how to wrap RSA-style probabilistic reasoning around "ground-truth" listener and speaker models, which are trained on human data, and how the resulting pragmatic models improve descriptions of visual scenes in a target-distractor setting. This general idea of grounding semantic meaning in a data-trained module and building a context-aware pragmatic RSA agent on top of it has been shown to boost performance also in a number of further applications, such as route descriptions (Fried et al. 2018), summarization and abstraction (Shen et al. 2019), image captioning (Cohn-Gordon et al. 2018, Nie et al. 2020), and machine translation (Cohn-Gordon & Goodman 2019).

7 Conclusion

As the title of the current work promises, our aim has been a practical introduction to the RSA modeling framework: what would the would-be modeler need to design their own computational cognitive model of probabilistic language understanding? We began with a high-level overview of the modeling framework, which treats language understanding as a process of probabilistic, recursive social reasoning; we also stressed how the formalization offers an articulated implementation of Gricean pragmatics. With an understanding of Vanilla RSA from Frank & Goodman (2012), which was used to model communication behavior in simple coordinated reference games, we then expanded our sights to a broader range of language phenomena, taking note of the technology used to capture them. We showed how expanding the empirical coverage of RSA often involves taking into

account additional sources of uncertainty: uncertainty about the semantics or lexicon, uncertainty about the QUD, uncertainty about the appropriate context, uncertainty about the epistemic states of conversational participants, and uncertainty about the speaker's utility calculus.

After considering the practical and theoretical benefit of thinking through models both "in math" and "in code", we then shifted our focus to issues that commonly arise in model development: how to set parameters and how to link model predictions to the empirical data they are meant to model. While these considerations occupy a significant portion of the modeler's time, they rarely feature prominently in model write-ups, which means that best practices can be hard to identify without an exercise in trial-and-error.

Finally, we explored current limitations of the modeling framework, together with areas where the framework is being actively extended.

Although our focus has been on the modeler, we hope to have offered valuable insight into the framework, which ought to provide even passive consumers with the ability to better evaluate and appreciate the growing RSA literature.

References

Akhtar, Nameera, Malinda Carpenter & Michael Tomasello. 1996. The role of discourse novelty in early word learning. *Child Development* 67(2). 635–645.

Andreas, Jacob & Dan Klein. 2016. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, 1173–1182. Austin, TX: Association for Computational Linguistics.

Augurzky, Petra, Michael Franke & Rolf Ulrich. 2019. Gricean expectations in online sentence comprehension: an ERP study on the processing of scalar inferences. *Cognitive Science* 43(8).

Benz, Anton & Robert van Rooij. 2007. Optimal assertions and what they implicate. *Topoi* 26. 63–78. \https://doi.org/doi:10.1007/s11245-006-9007-3.

Bergen, Leon, Roger Levy & Noah D. Goodman. 2016. Pragmatic reasoning through semantic inference. *Semantics and Pragmatics* 9(20).

Bohn, Manuel, Michael Henry Tessler & Michael C Frank. 2019a. Integrating common ground and informativeness in pragmatic word learning. In *Proceedings of the 41st annual meeting of the Cognitive Science Society*, .

Bohn, Manuel, Michael Henry Tessler, Megan Merrick & Michael C Frank. 2019b. Predicting pragmatic cue integration in adults' and children's inferences about novel word meanings .

Brochhagen, Thomas, Michael Franke & Robert van Rooij. 2018. Co-evolution of lexical meaning and pragmatic use. *Cognitive Science* 42(8). 2757–2789.
Brown, Penelope & Stephen C Levinson. 1987. *Politeness: Some universals in language usage*. Cambridge university press.
Carcassi, Fausto. 2020. *The cultural evolution of scalar categorization*: University of Edinburgh dissertation.
Carcassi, Fausto, Marieke Schwoustra & Simon Kirby. 2019. The evolution of adjectival monotonicity. In M. Teresa Espinal, Elena Castroviejo, Manuel Leonetti, Louise McNally & Cristina Real-Puigdollers (eds.), *Proceedings of Sinn und Bedeutung 23*, vol. 1, 219–230.
Chemla, Emmanuel & Raij Singh. 2014. Remarks on the experimental turn in the study of scalar implicature (part i & ii). *Language and Linguistics Compass* 8(9). 373–386, 387–399.
Chierchia, Gennaro, Danny Fox & Benjamin Spector. 2012. Scalar implicature as a grammatical phenomenon. In Klaus von Heusinger, Claudia Maienborn & Paul Portner (eds.), *Semantics: An international handbook of natural language meaning*, vol. 3, chap. 87, 2297–2331. Berlin: Mouton de Gruyter.
Chierchia, Gennaro & Sally McConnell-Ginet. 2000. *Meaning and grammar: An introduction to semantics.* Cambridge, MA: MIT Press.
Clark, Herbert H & Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition* 22(1). 1–39.
Cohn-Gordon, Reuben & Noah D. Goodman. 2019. Lost in machine translation: A method to reduce meaning loss. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies*, vol. 1, 437–441. Minneapolis, Minnesota: Association for Computational Linguistics.
Cohn-Gordon, Reuben, Noah D. Goodman & Christopher Potts. 2018. Pragmatically informative image captioning with character-level inference. In *Proceedings of the 2018 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies*, 439–443. Stroudsburg, PA: Association for Computational Linguistics.
Cohn-Gordon, Reuben, Noah D. Goodman & Christopher Potts. 2019. An incremental iterated response model of pragmatics. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, 81–90.
Crain, Stephen & Cecile McKee. 1985. The acquisition of structural restrictions on anaphora. In *Proceedings of NELS*, vol. 15, 94–110.
Crain, Stephen & Rosalind Thornton. 1998. *Investigations in Universal Grammar: A guide to research on the acquisition of syntax and semantics*. Cambridge, MA: MIT Press.

van Deemter, Kees, Ielka van der Sluis & Albert Gatt. 2006. Building a semantically transparent corpus for the generation of referring expressions. In *Proceedings of the fourth international natural language generation conference*, 130–132. Sydney, Australia: Association for Computational Linguistics.

Degen, Judith & Noah D Goodman. 2014. Lost your marbles? the puzzle of dependent measures in experimental pragmatics. In *Proceedings of the 36th annual conference of the Cognitive Science Society*, 397–402.

Degen, Judith, Robert X.D. Hawkins, Caroline Graf, Elisa Kreiss & Noah D. Goodman. 2020. When redundancy is useful: A Bayesian approach to 'overinformative' referring expressions. *Psychological Review* 127(4). 591–621.

Degen, Judith, Michael Henry Tessler & Noah D Goodman. 2015. Wonky worlds: Listeners revise world knowledge when utterances are odd. In *Proceedings of the 37th annual meeting of the Cognitive Science Society*, .

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

Frank, Michael C, Andres Gomez Emilsson, Benjamin Peloquin, Noah D Goodman & Christopher Potts. 2016. Rational speech act models of pragmatic reasoning in reference games .

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

Frank, Michael C & Noah D Goodman. 2014. Inferring word meanings by assuming that speakers are informative. *Cognitive psychology* 75. 80–96.

Franke, Michael. 2014a. Pragmatic reasoning about unawareness. *Erkenntnis* 79(4). 729–767. \https://doi.org/10.1007/s10670-013-9464-1.

Franke, Michael. 2014b. Typical use of quantifiers: A probabilistic speaker model. In Paul Bello, Marcello Guarini, Marjorie McShane & Brian Scassellati (eds.), *Proceedings of cogsci*, 487–492. Austin, TX: Cognitive Science Society.

Franke, Michael. 2016. Task types, link functions & probabilistic modeling in experimental pragmatics. In Fabienne Salfner & Uli Sauerland (eds.), *Proceedings of trends in experimental pragmatics*, 56–63.

Franke, Michael & Leon Bergen. 2020. Theory-driven statistical modeling for semantics & pragmatics: A case study on grammatically generated implicature readings. *Language* 96(2). e77–e96.

Franke, Michael, Fabian Dablander, Anthea Schöller, Erin Bennett, Judith Degen, Michael Henry Tessler, Justine T Kao & Noah D Goodman. 2016. What does the crowd believe? a hierarchical approach to estimating subjective beliefs from empirical data. In *Proceedings of the thirty-eighth annual conference of the Cognitive Science Society*, .

Franke, Michael & Judith Degen. 2016. Reasoning in reference games: Individual- vs. population-level probabilistic modeling. *PLoS ONE* 11(5). e0154854.

Franke, Michael & Gerhard Jäger. 2016. Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaf* 35. 3–44.

Fried, Daniel, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein & Trevor Darrell. 2018. Speaker-follower models for vision-and-language navigation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi & R. Garnett (eds.), *Advances in neural information processing systems*, vol. 31, 3314–3325. Curran Associates, Inc.

Gelman, Andrew, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari & Donald B Rubin. 2013. *Bayesian data analysis*. CRC press.

Geurts, Bart. 2010. *Quantity implicatures*. Cambridge, UK: Cambridge University Press.

Goodman, Noah D. & Michael C. Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences* 20(11). 818–829.

Goodman, Noah D. & Daniel Lassiter. 2015. Probabilistic semantics and pragmastics: Uncertainty in language and thought. In S. Lappin & C. Fox (eds.), *The handbook of contemporary semantic theory*, Wiley-Blackwell 2nd edn.

Goodman, Noah D. & Andreas Stuhlmüller. 2013. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science* 5. 173–184.

Goodman, Noah D. & Andreas Stuhlmüller. 2014. The design and implementation of probabilistic programming languages. \http://dippl.org.

Grice, Paul Herbert. 1975. Logic and conversation. In Peter Cole & Jerry L. Morgan (eds.), *Syntax and semantics, vol. 3, speech acts*, 41–58. New York: Academic Press.

Griffiths, Thomas L., Falk Lieder & Noah D. Goodman. 2015. Rational Use of Cognitive Resources: Levels of Analysis Between the Computational and the Algorithmic. *Topics in Cognitive Science* 7. 217–229.

Hawkins, Robert D, Michael C Frank & Noah D Goodman. 2020. Characterizing the dynamics of learning in repeated reference games. *Cognitive Science* 44(6). e12845.

Hawkins, Robert XD, Michael C. Frank & Noah D Goodman. 2017. Convention-formation in iterated reference games. In *Proceedings of the thirty-ninth annual conference of the Cognitive Science Society*, .

Herbstritt, Michele & Michael Franke. 2019. Complex probability expressions & high-order uncertainty: compositional semantics, probabilistic pragmatics & experimental data. *Cognition* 186. 50–71.

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

Hurford, James R. 1974. Exclusive or inclusive disjunction. *Foundations of Language* 11(3). 409–411.

Jasbi, Msoud, Brandon Waldon & Judith Degen. 2019. Linking hypothesis and number of response options modulate inferred scalar implicature rate. *Frontiers in Psychology* 10. 189.

Jeffreys, Harold. 1961. *Theory of probability*. Oxford: Oxford University Press 3rd edn.

Kao, Justine T, Leon Bergen & Noah D. Goodman. 2014a. Formalizing the pragmatics of metaphor understanding. In *Proceedings of the thirty-sixth annual conference of the Cognitive Science Society*, 719–724.

Kao, Justine T & Noah D. Goodman. 2015. Let's talk (ironically) about the weather: Modeling verbal irony. In *Proceedings of the thirty-seventh annual conference of the Cognitive Science Society*, .

Kao, Justine T, Jean Y Wu, Leon Bergen & Noah D. Goodman. 2014b. Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences of the United States of America* 111. 12002–12007.

Kass, Robert E. & Adrian E. Raftery. 1995. Bayes factors. *Journal of the American Statistical Association* 90(430). 773–795.

Katzir, Roni. 2007. Structurally-defined alternatives. *Linguistics and Philosophy* 30(6). 669–690.

Kruschke, John. 2014. *Doing bayesian data analysis: A tutorial with r, jags, and stan*. Academic Press.

Lambert, Ben. 2018. *A student's guide to bayesian statistics*. Sage Publications.

Lassiter, Daniel & Noah D. Goodman. 2013. Context, scale structure, and statistics in the interpretation of positive-form adjectives. In Todd Snider (ed.), *Proceedings of SALT 23*, 587–610. CLC Publications.

Lassiter, Daniel & Noah D Goodman. 2017. Adjectival vagueness in a bayesian model of interpretation. *Synthese* 194(10). 3801–3836.

Lee, Michael D & Eric-Jan Wagenmakers. 2014. *Bayesian cognitive modeling: A practical course*. Cambridge university press.

Lieder, Falk & Thomas L Griffiths. 2020. Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences* 43.

Lund, Gunnar, Rebecca Jarvis & Gregory Scontras. 2019. The pragmatics of semantic change: Modeling the progressive-to-imperfective shift. In *Proceedings of Sinn und Bedeutung 23*, 73–90.

Markman, Ellen M & Gwyn F Wachtel. 1988. Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive psychology* 20(2). 121–157.

Marr, David. 1982. *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: W. H. Freeman and Company.

Monroe, Will, Robert X. D. Hawkins, Noah D. Goodman & Christopher Potts. 2017. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics* 5. 325–338.

Monroe, Will & Christopher Potts. 2015. Learning in the Rational Speech Acts model. In *Proceedings of 20th Amsterdam Colloquium*, Amsterdam: ILLC.

Montague, Richard. 1973. The proper treatment of quantification in ordinary English. In K J J Hintikka, J M E Moravcsik & P Suppes (eds.), *Approaches to natural language*, vol. 49 Synthese Library, 221–242. Dordrecht: Reidel.

Nie, Allen, Reuben Cohn-Gordon & Christopher Potts. 2020. Pragmatic issue-sensitive image captioning. In *Findings of the association for computational linguistics: Emnlp 2020*, 1924–1938. Online: Association for Computational Linguistics.

Ohmer, Xenia, Peter König & Michael Franke. 2020. Reinforcement of semantic representations in pragmatic agents leads to the emergence of a mutual exclusivity bias. In *Proceedings of CogSci 42*, .

Potts, Christopher, Daniel Lassiter, Roger Levy & Michael C. Frank. 2016. Embedded implicatures as pragmatic inferences under compositional lexical uncertainty. *Journal of Semantics* 33(4). 755–802.

Qing, Ciyang & Michael Franke. 2014. Gradable adjectives, vagueness, and optimal language use: A speaker-oriented model. In Jessi Grieser, Todd Snider, Sarah D'Antonio & Mia Wiegand (eds.), *Proceedings of SALT 44*, 23–41. elanguage.net.

Qing, Ciyang & Michael Franke. 2015. Variations on a Bayesian theme: Comparing Bayesian models of referential reasoning. In H. Zeevat & H.-C. Schmitz (eds.), *Bayesian natural language semantics and pragmatics*, 201–220. Springer.

Roberts, Craige. 2012. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics* 5(6). 1–69.

Savinelli, K.J., Gregory Scontras & Lisa Pearl. 2017. Modeling scope ambiguity resolution as pragmatic inference: Formalizing differences in child and adult behavior. In *Proceedings of the thirty-ninth annual conference of the Cognitive Science Society*, 2064–3069.

Savinelli, K.J., Gregory Scontras & Lisa Pearl. 2018. Exactly two things to learn from modeling scope ambiguity resolution: Developmental continuity and numeral semantics. In *Proceedings of the 8th workshop on cognitive modeling and computational linguistics (CMCL 2018)*, 67–75.

Schöller, Anthea & Michael Franke. 2017. Semantic values as latent parameters: Testing a fixed threshold hypothesis for cardinal readings of *few* & *many*. *Linguistic Vanguard* 3(1).

Schulz, Katrin & Robert van Rooij. 2006. Pragmatic meaning and non-monotonic reasoning: The case of exhaustive interpretation. *Linguistics and Philosophy* 29. 205–250.

Schulze, Cornelia, Susanne Grassmann & Michael Tomasello. 2013. 3-year-old children make relevance inferences in indirect verbal communication. *Child development* 84(6). 2079–2093.

Schuster, Sebastian & Judith Degen. 2020. I know what you're probably going to say: Listener adaptation to variable use of uncertainty expressions. *Cognition* 203. 104285.

Scontras, Gregory & Noah D. Goodman. 2017. Resolving uncertainty in plural predication. *Cognition* 168. 294–311.

Scontras, Gregory, Michael Henry Tessler & Michael Franke. electronic. Probabilistic language understanding: An introduction to the Rational Speech Act framework. Retrieved from https://www.problang.org.

Sedivy, Julie C. 2007. Implicature during real time conversation: A view from language processing research. *Philosophy Compass* 2(3). 475–496.

Sedivy, Julie C., Michael K. Tanenhaus, Craig G. Chambers & Gergory N. Carlson. 1999. Achieving incremental semantic interpretation through contextual representation. *Cognition* 71. 109–147. \https://doi.org/10.1016/S0010-0277(99)00025-6.

Shen, Sheng, Daniel Fried, Jacob Andreas & Dan Klein. 2019. Pragmatically informative text generation. In *Proceedings of naacl*, .

Skyrms, Brian. 2010. *Signals: Evolution, learning, and information*. Oxford: Oxford University Press.

Spector, Benjamin. 2006. Scalar implicatures: Exhaustivity and gricean reasoning. In Maria Aloni, Alistair Butler & Paul Dekker (eds.), *Questions in dynamic semantics*, 229–254. Amsterdam, Singapore: Elsevier.

Tanenhaus, M.K., M.J. Spivey-Knowlton, K.M. Eberhard & J.C. Sedivy. 1995. Integration of visual and linguistic information in spoken language

comprehension. *Science* 268(5217). 1632–1634.
Tessler, Michael Henry & Noah D. Goodman. 2019. The language of generalization. *Psychological Review* 126. 395–436.
Tessler, Michael Henry, Michael Lopez-Brau & Noah D Goodman. 2017. Warm (for winter): Comparison class understanding in vague language. In *Proceedings of the 39th annual meeting of the Cognitive Science Society*, .
van Tiel, Bob, Michael Franke & Uli Sauerland. 2021. Probabilistic pragmatics explains gradience and focality in natural language quantification. Manuscript.
Werning, Markus & Erica Cosentino. 2017. The interaction of Bayesian pragmatics and lexical semantics in linguistic interpretation: Using event-related potentials to investigate hearers' probabilistic predictions. In Glenn Gunzelmann, Andrew Howes, Thora Tenbrink & Eddy Davelaar (eds.), *Proceedings of CogSci 39*, 3504–3509. Austin, TX: Cognitive Science Society.
Woensdregt, Marieke, Chris Cummins & Kenny Smith. 2020. A computational model of the cultural co-evolution of language and mindreading. *Synthese* .
Yoon, Erica J, Michael Henry Tessler, Noah D Goodman & Michael C Frank. 2016. Talking with tact: Polite language as a balance between kindness and informativity. In *Proceedings of the thirty-eighth annual conference of the Cognitive Science Society*, .
Yoon, Erica J, Michael Henry Tessler, Noah D Goodman & Michael C Frank. 2017. "i won't lie, it wasn't amazing": Modeling polite indirect speech. In *Proceedings of the thirty-ninth annual conference of the Cognitive Science Society*, .
Yoon, Erica J, Michael Henry Tessler, Noah D Goodman & Michael C Frank. 2020. Polite speech emerges from competing social goals. *Open Mind: Discoveries in Cognitive Science* \https://doi.org/10.31234/osf.io/67ne8. \psyarxiv.com/67ne8.
Zaslavsky, Noga, Jennifer Hu & Roger P Levy. 2020. A rate-distortion view of human pragmatic reasoning. *arXiv preprint arXiv:2005.06641* .