CellPress

Review

Pragmatic Language Interpretation as Probabilistic Inference

Noah D. Goodman[1,*] and Michael C. Frank[1]

Understanding language requires more than the use of fixed conventions and more than decoding combinatorial structure. Instead, comprehenders make exquisitely sensitive inferences about what utterances mean given their knowledge of the speaker, language, and context. Building on developments in game theory and probabilistic modeling, we describe the rational speech act (RSA) framework for pragmatic reasoning. RSA models provide a principled way to formalize inferences about meaning in context; they have been used to make successful quantitative predictions about human behavior in a variety of different tasks and situations, and they explain why complex phenomena, such as hyperbole and vagueness, occur. More generally, they provide a computational framework for integrating linguistic structure, world knowledge, and context in pragmatic language understanding.

'...one of my avowed aims is to see talking as a special case or variety of purposive, indeed rational, behavior' Grice ([1] p. 47).

Trends

Rational speech act (RSA) models provide a quantitative framework to capture intuitions about pragmatic reasoning in language understanding.

Extensions to RSA that allow for reasoning about the speaker (for instance, her goals and word usage) can capture many otherwise puzzling phenomena, including vagueness, embedded implicatures, hyperbole, irony, and metaphor.

The RSA framework can inform psycholinguistic processing experiments, linguistic theory, and scalable natural language processing models.

Understanding Language

Language is central to the successes of our species; with language, we can coordinate our actions, learn from each other, and convey our innermost thoughts. From sounds to syntax, natural languages provide structured methods of combining discrete materials to generate an infinite variety of sentences. Yet, this discrete combinatorics does not fully explain how speakers can use language so flexibly to achieve social goals. The interpretation of a particular utterance can itself be almost infinitely variable, depending on factors such as the identity of the speaker, the physical context of its use, and the previous discourse. While the systematization of structural features of language is one of the proudest accomplishments of cognitive science (e.g., [2–4]), its contextual flexibility (its pragmatics) has been stubbornly difficult to formalize.

Grice [1] presented an initial framework theory for pragmatic reasoning, positing that speakers are taken to be cooperative, choosing their utterances to convey particular meanings. Gricean listeners then attempt to infer the speaker's intended communicative goal, working backward from the form of the utterance. This goal inference framework for communication has been immensely influential (e.g., [5–8]). However, attempts to build on these ideas by providing a specific set of formal principles that allow the derivation of pragmatic inferences have met with difficulty.

For example, the core of Grice's proposal was a set of **conversational maxims** (see Glossary). Inferences about speakers' behavior relative to these maxims (be truthful, relevant, informative, and perspicuous) could lead to **implicatures** (inferences about their intended meaning).

[1]Department of Psychology, Stanford University, 450 Serra Mall, Stanford, CA 94305, USA

*Correspondence: ngoodman@stanford.edu (N.D. Goodman).

 http://dx.doi.org/10.1016/j.tics.2016.08.005

However, formalization of the Gricean notion of implicature using the maxims is difficult and many post-Gricean theories have instead proposed alternative sets of principles [6,8]. An important test of the difficulty of this theoretical project is that the burgeoning experimental psycholinguistic literature attempting to measure pragmatic inference has found these principles to be only modestly useful [9–11]. In addition, this kind of informal theory of pragmatics can make only directional, qualitative predictions with respect to experimental data that are typically graded and quantitative.

An alternative strand of Gricean thought has had more success in making contact with data. Grice's core insight was that language use is a form of rational action; thus, technical tools for reasoning about rational action should elucidate linguistic phenomena. Such a goal-directed view of language production has led to the development of engineering systems for natural language generation [12] that have in turn been applied as theories of human language production (e.g., [13]). Concurrently, the tools of game theory, which allow for the characterization of rational actions with respect to defined utilities, have provided a vocabulary for formal descriptions of pragmatic phenomena (e.g., [14,15]). The recent work we focus on here builds on these developments, combining them with a more detailed view of cognition that arises from the Bayesian cognitive modeling tradition.

Probabilistic, or Bayesian, models have been at the core of a set of recent attempts to understanding the interplay between structured representations and graded or statistical information [16]. These models have been an important tool for understanding nonlinguistic varieties of rational action, integrating belief understanding with action planning [17]. A critical feature of these models is that they use the probability calculus to describe inferences under uncertainty. Within formal models of pragmatics, this uncertainty stems from a variety of sources, including uncertainty about speakers' goals and beliefs, uncertainty about the discourse and broader context, and even uncertainty about the meanings of words.

In the remainder of this paper, we describe the probabilistic approach to pragmatics. We begin by presenting the **rational speech act (RSA) model**, and the growing body of empirical data supporting its utility in explaining pragmatic reasoning. We next discuss extensions to RSA that allow it to be applied to nonliteral uses of language, such as hyperbole, irony, and metaphor, to cases of vagueness and ambiguity, and to complex interactions between pragmatics and compositional syntax and/or semantics. We close by considering the broader applications of, and challenges for, probabilistic pragmatics models.

A 'Rational Speech Act' Model

The RSA model implements a social cognition approach to utterance understanding. At its core, it captures the idea (due to Grice, David Lewis, and others) that speakers are assumed to produce utterances to be helpful yet parsimonious, relative to some particular topic or goal. Listeners then understand utterances by inferring what such a helpful speaker must have meant, given what she said.* The first of these basic assumptions is formalized by viewing the speaker as a utility-maximizing agent (where the effort of language production is costly, but communicating information is beneficial). The listener then updates his beliefs via Bayesian inference.

The pragmatic listener infers the state of the world, w, using Bayes' rule, given the observation that the speaker chose a particular utterance, u (Equation 1):

$$P_L(w|u) \propto P_S(u|w)P(w). \quad [1]$$

Glossary

Conversational maxims: a set of principles described by Grice [1] as a theory of how listeners reason about speakers' intended meaning to arrive at pragmatic implicatures.
Implicature: an inference about the meaning of an utterance in context that goes beyond its literal semantics. Implicatures are typically cancellable in that they can be contradicted, as in 'Some of the students passed the exam; indeed, all of them did'.
Rational speech act model (RSA): a class of probabilistic model that assumes that language comprehension in context arises via a process of recursive reasoning about what speakers would have said, given a set of communicative goals.
Scalar implicature: an implicature that arises when a speaker did not use a stronger alternative term, leading to narrowing of the interpretation. For instance, hearing 'some' usually leads to a scalar implicature that 'not all'.
Social recursion: reasoning that involves two people, such as listener and speaker, thinking about each other: 'I think that you think that I think that...'. This may proceed to a finite depth, or continue *ad infinitum*.
Uncertain RSA models (uRSA): a specific extension of RSA models that allow for joint inferences about both the speaker's intended meaning and other aspects of the interaction, such as the topic, the context, or even the meanings of particular words.

*For clarity throughout, we use a female pronoun for Alice, the speaker, and a male pronoun for Bob, the listener.

The key assumption he must make is that the speaker is approximately rational; that is, that she has chosen her utterances in proportion to the utility she expects to gain (Equation 2):

$$P_S(u|w) \propto \exp(\alpha U(u;w)). \quad [2]$$

The speaker chooses u from a set of alternative utterances (see Outstanding Questions). The parameter α captures the extent to which the speaker maximizes her utility: how rational she will be. The basic speaker utility used in RSA captures the social benefit of providing epistemic help to a listener (Equation 3):

$$U(u;w) = \log P_{\text{Lit}}(w|u). \quad [3]$$

Eq. (3) measures how certain the listener becomes about the intended world after hearing the utterance; to avoid an infinite recursion and provide an entry point for conventional (semantic) meaning, the speaker is assumed to consider a simpler listener, the 'literal listener' P_{Lit}. The literal listener again updates his beliefs in accord with Bayesian inference, under the assumption that the literal meaning of the utterance is true (Equation 4):

$$P_{\text{Lit}}(w|u) \propto \delta_{[\![u]\!](w)} P(w). \quad [4]$$

This definition of the literal listener requires a semantic denotation for each sentence, $[\![u]\!]$, in which a sentence has the value true or false when applied to a particular state of affairs, w (the 'world'). This denotation is how conventional meaning enters the pragmatic reasoning process and it connects RSA to work in lexical and compositional semantics [18,19].

Consider the scenario in Figure 1, in which the speaker and the listener share a world of three faces: one with hat and glasses (HG), one with glasses only (G), and one with neither (N); one of these (known to the speaker but not the listener) is the 'friend'. The speaker says 'my friend has glasses', presupposing that there is a single friend. Experimental participants, who know that the only alternative utterance was 'my friend has a hat', tend to share the intuition that this sentence refers to G and not HG or N [20,21]. While real-world language has many more utterances available (e.g., 'my friend has glasses, but no hat'), this restricted scenario serves to illustrate the underlying dynamics of pragmatic reasoning.

Under RSA, listener L reasons about S (a simplified internal representation of the speaker), who in turn reasons about Lit (a yet more simplified internal model of the listener). Lit updates his beliefs based on a straightforward denotation: 'glasses' applies to both G and HG, while 'hat' applies only to HG. Thus $P_{\text{Lit}}(w|''\text{hat}'')$ places all probability on the friend being HG, while $P_{\text{Lit}}(w|''\text{glasses}'')$ places equal probability on G and HG (see innermost thought bubbles in Figure 1). Thus, the speaker S who intends to communicate that HG is the friend will tend to choose the more informative 'hat'; but if she intends to communicate that G is the friend, she will use 'glasses'. Finally, upon hearing 'glasses', the listener L infers that this likely refers to G (reflecting the counterfactual that, if S had been talking about HG, she would have said 'hat' instead).

In the simplest RSA model, as illustrated above, the speaker values providing epistemic help (information) to the listener. However, the model can also be extended to create a more sophisticated speaker who is uncertain about the world state, who avoids costly utterances, or who aims to provide relevant information (Box 1). Connections to other theoretical approaches and aspects of language then become straightforward. For instance, by modifying the speaker's utility function, we can model the notion of topic-relevant information, which connects to linguistic ideas about the 'question under discussion' [22]. As a second example, RSA can be combined with the noisy channel approach to language comprehension [23] to explain the communicative use of sentence fragments and prosodic stress [24].

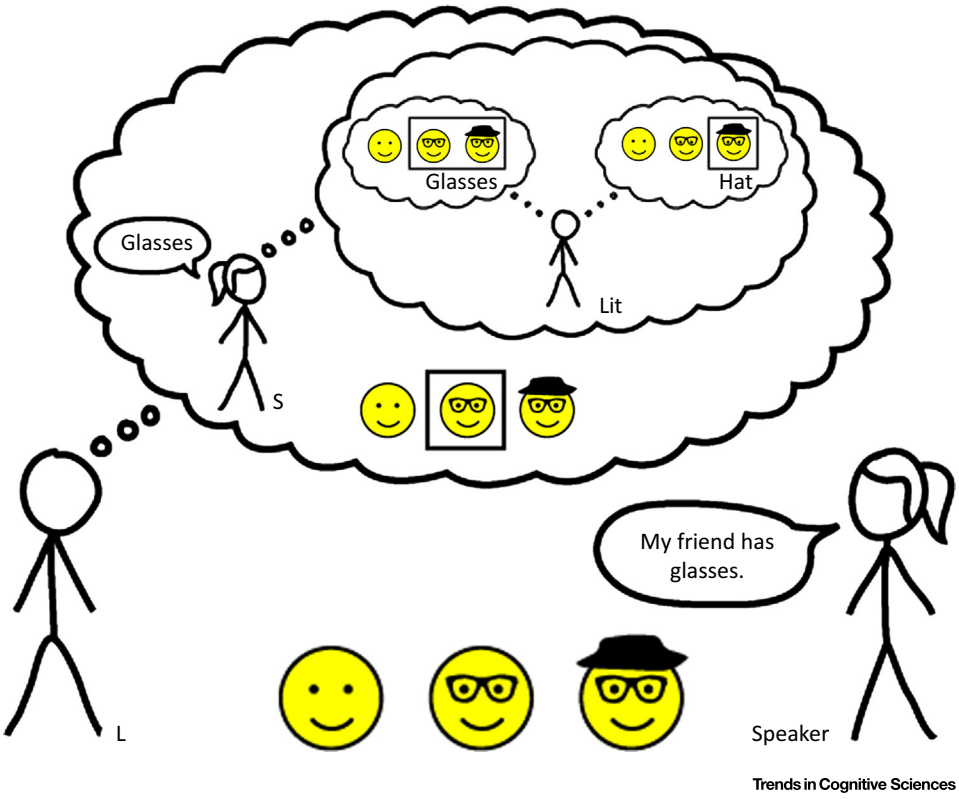

Figure 1. Application of Rational Speech Act-Style Reasoning to a Signaling Game. The three faces along the bottom show the signaling game context. Agents are depicted as reasoning recursively about one another's beliefs: listener L reasons about an internal representation of a speaker S, who in turn is modeled as reasoning about a simplified literal listener, Lit. Boxes around targets in the reference game denote interpretations available to a particular agent.

In sum, RSA models replace Grice's maxims with a single, utility-theoretic version of the cooperative principle [25]. This formulation is based on utilities that can reflect the communicative and social priorities of a complex, real-world agent.

Empirical Support for RSA

The example shown above in Figure 1 is an instance of a signaling game of the type initially introduced by Lewis [26]. Such games are a valuable tool for exploring pragmatic inferences in context, and experiments testing the RSA framework have used games of this type to make quantitative measurements of a variety of different inferences. For example, one paper [27] used a one-shot, web-based paradigm to present participants with geometric shapes in a variety of different configurations. Using a betting paradigm (participants were asked to distribute US$100 between response options), a set of experiments collected separate judgments about what a speaker would say, what a listener would interpret, and the baseline expectations for reference [corresponding to the prior P(w)]. The RSA model showed a tight fit to listeners' aggregate judgments when combined with empirical measurements of the prior distribution: P_S and P_L models correlated strongly with participants' average bets on what to say and how to interpret, respectively.

Although in this initial work RSA was used to simulate the behavior of both speakers and listeners, most subsequent work has focused on the behavior of listeners alone. This work

Box 1. Refinements to the Speaker's Utility

The notion of the speaker's utility (what is rewarding for a speaker) is central to the RSA approach. The basic RSA model captures the speaker's need to be informative to a listener (Equation I):

$$U(u;w) = \log P_{\text{Lit}}(w|u). \quad [\text{I}]$$

Different utilities lead to different kinds of speaker, which in turn lead to different interpretations by the pragmatic listener. Several utility refinements (and their combinations) have been considered in recent work:

- Utterance cost: to capture a tendency of speakers to be parsimonious we can simply add a cost term (Equation II):

  $$U(u;w) = \log P_{\text{Lit}}(w|u) + \text{cost}(u) \quad [\text{II}]$$

  The cost may reflect actual production cost (such as number of words) or proxies, such as word frequency. This extension yields effects similar to Grice's maxim of manner [47].

- Speaker uncertainty: when the speaker does not have full knowledge of the world she should choose an utterance according to expected utility (Equation III):

  $$U(u;k) = \mathbb{E}_{P(w|k)}[U(u;w), \quad [\text{III}]$$

  where k summarizes the speaker's knowledge or observations. This extension correctly predicts interactions between a speaker's knowledge and a listener's interpretations [38].

- Topic relevance: although it may be highly informative to provide detailed descriptions, such detail is not always relevant. Relevance can be captured by introducing a topic of conversation [22], sometimes known as a 'Under Discussion' [22] and adjusting the epistemic utility to reflect only information about this topic (Equation IV):

  $$U(u;w,t) = \log \sum_{w' \text{s.t.} t(w')=t(w)} P_{\text{Lit}}(w'|u). \quad [\text{IV}]$$

  Here, the topic is encoded in a function t that takes a complete world and yields some subset or summary; for instance, in the case of hyperbole [40], t can pick out only the speaker's affect, dropping objective states.

- Other social goals: language is often used not just to inform, but also to flirt, insult, comfort, and pursue myriad other social goals. For example, non-informational utilities, such as utility directed toward kindness, can produce behaviors that appear polite [72].

Box 2. Producing Referring Expressions

RSA stands for the 'rational speech act' model, indicating that listeners idealize speakers as rational. Are speakers in fact rational in a meaningful way? If so, how can this conclusion be integrated with the large body of evidence indicating that speakers are egocentric, error prone, and subject to idiosyncratic production preferences [67–69]?

Although our initial studies collected judgments about language production in extremely restricted tasks [27], most recent work using the RSA model has focused on modeling listeners' judgments, rather than speakers' productions. One reason for this choice is that often the most interesting pragmatic inferences come about when speakers are not maximally informative. For example, in the signaling game shown in Figure 1, helpful speakers will often overspecify and say 'glasses and no hat' [70]. However, this unnecessarily redundant utterance may in fact be a reasonable response to uncertainty about whether a conversational partner will in fact draw the desired implicature. More generally, speakers' production choices are a promising area for future research using RSA models with a broader range of utility functions (Box 1) and that incorporate various sources of potential miscommunication (a topic of ongoing research).

Nevertheless, it is clear that, in their natural behavior, speakers make production decisions under time pressure and a variety of cognitive demands [71]. Integrating these demands with the predictions of utility-theoretic models should be an important challenge for future work.

follows the idea that RSA captures listeners' (perhaps optimistic) assumptions about the rational behavior of speakers. Thus, RSA is 'rational' in the sense of assuming that speakers are rational; a separate question is how rational speakers in fact are (Box 2). In addition, though most research using RSA models has focused on mature language comprehenders, these models have also been the inspiration for a body of developmental work (Box 3).

A variety of other work has replicated and extended the initial findings using similar signaling-game paradigms. A tight replication of the initial results [28] reproduced the basic findings and

Box 3. Rational Speech Act (RSA) and Children's Developing Pragmatic Competence

From a very early age, children are oriented toward communication, understanding the function of language for information transfer and repurposing their limited linguistic means to achieve a variety of ends [60,61]. In light of this general orientation, the literature on pragmatic development specifically has been puzzling: older children reliably fail to make **scalar implicatures** under a range of circumstances [62]. In one striking example, most 5-year-olds endorsed the statement that 'some of the horses jumped over the fence' even when all three (out of three) of a set of horses had made the jump [63]. RSA-style models can provide a framework for thinking about this disconnect between early communicative successes and later pragmatic failures.

Recently, theorists proposed that children's apparent difficulties with pragmatic implicatures may have resulted from their inadequate knowledge of relations between lexical alternatives rather than difficulty with pragmatic computations more generally [64]. Congruent with this idea, 3-year-old children show signs of successful implicature computations in the kinds of signaling game shown in Figure 1, where the referential alternatives are all simple objects that are visible in the scene [21]. In addition, children in the same age range were able to use an implicature to guess the meaning of a novel word [65] or a novel context [66], showing the kind of inferential flexibility posited in RSA accounts. These findings support the idea that even young children are able to make flexible pragmatic inferences, and are consistent with the application of RSA-style reasoning, albeit with limits on the available semantic alternatives. However, future research will be required to test whether RSA (or some capacity-limited modification) could make quantitative predictions about pragmatic development.

explored a set of variants to the initial RSA utility function. And another study [29] found that RSA predicted judgments in a communication game using more complex spatial language stimuli, albeit with somewhat noisier fits. Thus, RSA with an epistemic utility can predict judgments in simple signaling games across variations in both participant sample and stimulus.

One question raised by this initial work was the level of **social recursion** that best fits human performance. The presentation of RSA given above is stated in terms of a minimal recursion (a listener reasons about speaker, who, in turn, reasons about a literal listener) but greater depths of reasoning are in principle possible. The evidence is mixed on whether deeper levels of recursion are commonly seen in language comprehension. In a variety of experiments exploring this issue, participants tended to show chance-level performance for signaling systems that required deeper levels of recursion to find unique interpretations [20,30,31]. However, more recent studies [32] showed some evidence of deeper recursion for a subpopulation of participants (approximately 15%) in a more complex paradigm, consistent with work on competitive economic games where deeper recursions are sometimes found [33]. This heterogeneity, and its dependence on individual and contextual differences, is an interesting topic for future work.

Several other studies have tested RSA with more elaborated utility functions (Box 1). For instance, a speaker might be expected to produce a less informative utterance when the more informative one is harder to say. This tendency can be formalized by including a cost term in the speaker's utility; with this modification, RSA predicts the impact of production costs on listeners' interpretations. Work exploring this extension [34] showed that participants in a reference game are indeed sensitive to the cost of message choices: the effect of alternative possible messages on a listener's inferences is modulated by their cost, in dollars. Related studies [35] tested the effect of production difficulty by manipulating how quickly the speaker could type on an on-screen keyboard; participants' interpretations reflected this difficulty as predicted. Additional work has used proxies for production cost, such as number of words and their frequencies in explorations of phenomena such as negation [36] or the choice of noun used to refer to an object [37].

Finally, in addition to ad-hoc signaling systems, RSA provides a way to describe reasoning about classic linguistic implicatures. Perhaps the best studied of these is the scalar implicature that

'some of the letters had checks inside' implicates that not all did. One study [38] measured participants' judgments about the interpretations of quantifiers and number words in exactly this situation and found that these judgments were well predicted by RSA. In addition, a critical feature of this study was the inclusion of an epistemic manipulation (e.g., whether all of the letters had already been opened). By using expected informativity (Box 1) to account for the speaker's limited perceptual access, the model was able to predict differing patterns of listener judgments based on different levels of speaker uncertainty. These empirical findings are congruent with other recent demonstrations of the importance of epistemic reasoning in pragmatic implicature [34,39], and the theoretical account accords with other probabilistic treatments of scalar implicature (B. Russell, PhD thesis, Brown University, 2012). They also highlight the way in which the RSA framework provides a (non-modular) theory for interactions between language and non-linguistic cognition. We next turn to a variety of other extensions to the basic RSA model that explore additional interactions.

Uncertainty about the Speaker: Joint Reasoning

In the basic RSA model, the listener has a specific model in mind of how a speaker will behave. However, what should a listener do if he is not sure what speaker model is appropriate? Speakers can differ in knowledge, communicative goals, and many other aspects; these differences can lead a listener to arrive at different interpretations of the same utterance. Recent work has addressed this issue by positing a joint inference: what type of speaker am I interacting with and what is the world like, given the utterance I heard? Formally this **uncertain RSA** (or **uRSA**) framework requires only a small change (Equation 5):

$$P_L(w,s|u) \propto P_S(u|w,s)P(s)P(w), \quad [5]$$

where the new variable s parametrizes different speaker types. In practice, s can refer to any factor that might influence the speaker's behavior, including uncertainty about conversational topic, word meanings, background knowledge, or general discourse context. This modification allows uRSA to capture a wider variety of linguistic phenomena; intuitively, an uRSA listener is a more realistic cognitive agent than the RSA listener, who was restricted to the specifics of a particular context and goal. To illustrate this intuition, we provide three examples of phenomena captured by uRSA (but not by basic RSA): nonliteral language, vagueness, and embedded implicatures.

Nonliteral or figurative language (utterances that are easily interpreted but not 'actually true') poses a problem for nearly all formal models of language understanding. How can tropes, such as hyperbole, sarcasm, and metaphor, be interpreted, and why are they used? Under uRSA, these uses can be described as arising from uncertainty about the topic of conversation. If the speaker is expected to provide information relevant for a particular topic, the pragmatic listener will only update his beliefs along this topical dimension. Within uRSA, the interaction between uncertainty about the speaker's intended topic and her intended meaning about that topic can drive complex interpretations.

Hyperbolic utterances such as 'the electric kettle cost $1000' are a key case study [40]. In this example, the number $1000 can be interpreted as conveying information about the speaker's affect, not the actual price, in part because one thousand dollars is an implausibly high price for a kettle. As shown in Figure 2, the uRSA model captures this intuition by positing that the topic of the speaker's utterance may be the actual price of the kettle, the speaker's opinion about the price, or some combination of the two. Since the listener does not know the topic, he jointly infers it together with the likely true price of the kettle and the speaker's affect. When the uttered price is implausible, it becomes more likely that the speaker is aiming to convey her opinion, and using $1000 (a price that most people would find too high) to do so. In this way, the listener's joint inference can yield a nonstandard topic and, hence, a nonliteral interpretation.

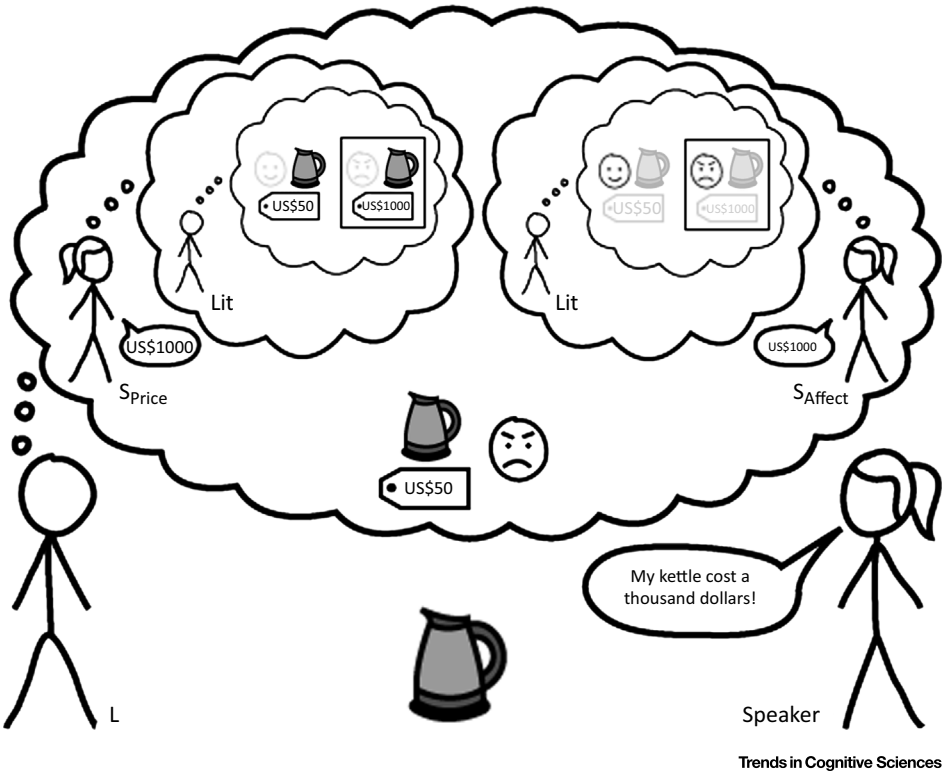

Figure 2. Uncertain Rational Speech Act-Style Reasoning Applied to Hyperbole. Listener L reasons jointly about the price of the item and the speaker's affect. In doing so, he considers two speakers, one who is primarily interested in conveying her affective response to the kettle, and one who is primarily interested in conveying the actual price. (The full model also considers speakers, not pictured, who wish to convey approximate price and combinations of these goals.) Each of these speakers is modeled as reasoning about a literal listener who interprets the utterance literally (indicated by the box selecting the 'US$1000' state), but focus on different aspects of the situation (price on the left and affect on the right).

By extending the space of affect to include both valence and arousal, the same model predicts verbal irony [41]. A similar approach has been suggested for simple metaphors [42], such as 'John is a shark'. Here, the potential topics include not affect, but features of the target, such as how vicious John is and how likely he is to swim underwater. In each of these cases of figurative language, the uRSA model accounts for almost all of the explainable variance in human interpretations, a striking result considering the complexity and subtlety of these phenomena.

Many linguistic descriptions, especially adjectives, are both context sensitive and vague. Providing precise definitions for words such as 'expensive' or 'tall' has been a persistent challenge for philosophers and semanticists [43]. uRSA models address this challenge by assuming that word meanings can differ between speakers and contexts, and and that these meanings themselves can be a subject for inference. In the case of scalar adjectives, such as 'tall', the uncertainty is over the threshold required: what height is required before an object counts as tall? Under the uRSA model, judgments about meaning take into account two conflicting pressures: on the one hand, a stricter threshold for tallness makes the term 'tall' more informative. For instance, 'Bob is tall' tells us a great deal if 'tall' requires a height greater than 8 ft. However, on the other hand, a stricter threshold makes such a sentence quite unlikely to be true *a priori*. By negotiating this balance between informativity and plausibility, uRSA accounts for three key phenomena of vague adjectives [44]: the inferred meaning depends on the class (tall for a tree versus tall for a person), there are borderline cases, and the interpretations

Box 4. Language Use and Language Change

The pragmatic processes described by RSA models occur in the moment of communication, but can have a set of effects that ripple out through language as a whole. The construal of an individual communication event can influence learning processes, which in turn can lead to systematic changes in word meanings [73]. Words that are too narrow in their denotation can be pragmatically extended [40], while words that are too broad can be narrowed via implicature. Over time, word meanings may converge to the appropriate level of ambiguity to enable efficient communication [74]. In this sense, in-the-moment pragmatic interpretation may bootstrap long-term language change.

The processes of change that promote efficient communication have been explored extensively within the iterated learning paradigm [75]. This framework can also be used to express the competing pressures of learnability and communication. When languages are selected only to be learnable, they often become degenerate, including only a single word [76]. However, when they include a countervailing pragmatic pressure, which can be modeled via RSA, expressive and compositional languages can emerge [77].

If pressures for efficient communication lead to language change, then these pressures should be visible in the lexicons of human languages. Indeed, a recent body of evidence suggests that the typological distribution of languages in particular semantic domains reflects the range of optimal communication systems (e.g., [78–80]). An important future direction is to understand whether this typological distribution is predicted to arise from iterated learning with a population of RSA-like language users.

Outstanding Questions

Social recursion: how deeply do human comprehenders reason about others' intentions? Is depth of recursion ('I think that you think that…') constant, or does it vary across situations?

Alternatives: how are alternative utterances computed? Do they depend on the language grammar? On situational factors?

Linguistic goals: how do 'Gricean' utilities (the drive to be informative yet succinct) relate to other social goals such as conveying affect or establishing relationships? How do cooperative and competitive goals mix in language use?

Dialogue: how can rational speech act (RSA) be used to model the evolution over the course of a conversation of a partner's utilities, possible goals, and the context more broadly?

Learning and language change: how do pragmatic language understanding and language learning interact? How and when does pragmatic language use lead to language change?

Algorithmic challenges: given the potential complexity of recursive pragmatic computations, how is language processed so quickly? How can RSA models be 'scaled up' for natural language processing tasks?

are subject to a *sorites* paradox (no single minimal increment in height will make you tall, but enough increments will). The processes of reasoning about meaning that are modeled by uRSA might even interact with learning processes to produce more long-lasting inferences about word meaning, leading to language change (Box 4).

Finally, this uRSA approach allows for progress on an important puzzle in recent discussions of pragmatic inference: embedded implicature [45,46]. Embedded implicatures occur when quantifiers are nested within one another, as in sentences such as 'Exactly one letter is connected with some of its circles'. In these cases, some experimental evidence suggests that participants access the interpretation that one letter is connected with some but not all of its circles, an interpretation that standard Gricean theories cannot generate [46]. Recent work [47] has replicated these interpretations in a series of large-scale experiments and confirmed that basic RSA models could not capture them. An implementation of uRSA that jointly infers word meanings and world state [47,48] showed a good fit to the overall pattern of data, however sentence meanings in this model are built by composing uncertain word meanings, showing how uRSA is a fruitful way to incorporate pragmatic reasoning into compositional semantic systems.

Concluding Remarks and Future Directions

Context dependence is one of the core features of natural language. Yet, because of the informal nature of theorizing about this context dependence, pragmatics has often been treated as a theoretical 'wastebasket', in which unexplained phenomena are hidden [49]. Countering this trend, new formal theories of pragmatics make quantitative predictions about a variety of phenomena that have previously been considered too difficult to operationalize. These include implicature, vagueness, non-literal language, and the myriad other cases where linguistic meaning is changed by context.

The key tool in this work is the Rational Speech Act framework, which builds upon and synthesizes a number of formal traditions in the study of human inference, from game theory to models of human reasoning. The RSA approach also builds on existing work on semantic representation, using a compositional semantics à la Montague [19], and contributes back to semantics, providing a specific mechanism by which underspecified meanings become precise in context. Rather than formalizing only a single hypothesis about pragmatic language understanding, RSA provides a framework in which many variations can be explored. Varying

assumptions about the speaker's utility (Box 1) and listener's uncertainty (see 'Uncertainty about the Speaker'), for instance, yields a spectrum of hypotheses that can be evaluated against quantitative experimental data.

While it has been successful in many recent cases, it may emerge that the RSA approach is not able to capture some aspects of language understanding, either because the foundational, Bayesian, tools it relies on are inadequate [50], or because pragmatic effects arise from sources not easily incorporated into RSA. Optimistically, however, RSA can be combined with other approaches when needed. For instance, the alternative utterances in RSA can be restricted [47] using previously proposed grammatical mechanisms [51]. In addition, increasingly, methods in machine learning have been used to supplement RSA with powerful learning mechanisms [52]. This cross-fertilization is among the most encouraging outcomes of work on RSA.

The RSA framework is a computational-level description of the language user's competence, in Marr's sense [53]. There are many possible ways a cognitive agent could implement RSA at the algorithmic level, and it is unclear which might best match the speed and competence of human language understanding and production. These alternatives must further be evaluated for their ability to explain the processing signatures of language comprehension, such as reaction times and eye gaze [36,54]. Yet, even as a computational-level framework, RSA inspires different intuitions about processing compared with previous theories. For example, RSA-style reasoning makes pragmatic inferences a fundamental part of language comprehension, in which the ultimate goal of all interpretation is to settle on the intended meaning, given both the literal semantics of the utterance and the broader pragmatic context. This framing contrasts with Gricean analyses, in which pragmatics enters when the violation of a maxim leads to reasoning to 'repair' the interpretation and correspondingly slower processing, a view that has been challenged both theoretically and empirically (e.g., [8,55]).

Future extensions of RSA will likely include worlds with richer structure; a thorough and practical theory of pragmatic alternatives; more sophisticated discourses that unfold over many utterances; and utility structures that better take into account the complexities of social interaction. On the practical side, computing the predictions of RSA models can become prohibitive when the number of world states or utterances grows large. Further development of algorithms to implement RSA is needed. These developments may go together with new algorithms for learning aspects of the underlying semantics, which will open new applications for the RSA approach in computational linguistics and artificial intelligence [52,56–59].

The work outlined in this review represents steps toward a comprehensive, formal theory of language understanding in context. Although further work will be required, RSA models and their uRSA extensions have proven to be useful tools for explaining both qualitative and quantitative empirical data across a range of tasks and contexts. Language is central to the human experience. We hope that our work sheds light on how its structure and systematicity can still give rise to such an astonishingly flexible communication system.

Acknowledgments

We gratefully acknowledge funding from a James S. McDonnell Foundation Scholar Award to Goodman, ONR Grants N00014-13-1-0788 and N00014-13-1-0287 to N.D.G. and M.C.F., and NSF BCS Grant 1456077 to M.C.F. Thanks to Allison Kraus for assistance with the illustrations in Figures 1 and 2.

References

1. Grice, H.P. (1975) Logic and conversation. In *Syntax and Semantics* (Vol. 3) Cole, P. and Morgan, J.,eds., pp. 41–58, Academic Press
2. Chomsky, N. (1965) *Syntactic Structures,* Walter de Gruyter
3. Jackendo, R.S. (2002) *Foundations of Language: Brain, Meaning, Grammar, Evolution,* Oxford University Press
4. Goldberg, A.E. (2003) Constructions: a new theoretical approach to language. *Trends Cogn. Sci.* 7, 219–224

5. Horn, L.R. (1984) Toward a new taxonomy for pragmatic inference: Q based and R-based implicature. In *Meaning, Form, and Use in Context: Linguistic Applications* (Schiffrin, D., ed.), pp. 11–42, Georgetown University Press
6. Sperber, D. and Wilson, D. (1986) *Relevance: Communication and Cognition,* Blackwell
7. Clark, H. (1996) *Using Language,* Cambridge University Press
8. Levinson, S. (2000) *Presumptive Meanings: The Theory of Generalized Conversational Implicature,* MIT Press
9. Breheny, R. *et al.* (2006) Are generalised scalar implicatures generated by default? An on-line investigation into the role of context in generating pragmatic inferences. *Cognition* 100, 434–463
10. Huang, Y.T. and Snedeker, J. (2009) Online interpretation of scalar quantifiers: insight into the semantics–pragmatics interface. *Cogn. Psychol.* 58, 376–415
11. Noveck, I.A. and Reboul, A. (2008) Experimental pragmatics: a Gricean turn in the study of language. *Trends Cogn. Sci.* 12, 425–431
12. Dale, R. and Reiter, E. (1995) Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cogn. Sci.* 19, 233–263
13. Viethen, J. and Dale, R. (2006) Algorithms for generating referring expressions: do they do what people do? In *Proceedings of the Fourth International Natural Language Generation Conference*, pp. 63–70, Association for Computational Linguistics
14. Benz, A. *et al.* (2006) An introduction to game theory for linguists. In *Game Theory and Pragmatics*, pp. 1–82, Springer
15. Jäger, G. (2008) Applications of game theory in linguistics. *Lang. Linguist. Compass* 2, 406–421
16. Tenenbaum, J. *et al.* (2011) How to grow a mind: statistics, structure, and abstraction. *Science* 331, 1279
17. Baker, C.L. *et al.* (2009) Action understanding as inverse planning. *Cognition* 113, 329–349
18. Heim, I. and Kratzer, A. (1998) *Semantics in Generative Grammar* (Vol. 13), Blackwell
19. Dowty, D.R. *et al.* (2012) *Introduction to Montague Semantics* (Vol. 11), Springer Science & Business Media
20. Stiller, A. *et al.* (2011) Ad-hoc scalar implicature in adults and children. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*, pp. 2134–2139, Cognitive Science Society
21. Stiller, A.J. *et al.* (2015) Ad-hoc implicature in preschool children. *Lang. Learn. Dev.* 11, 176–190
22. Roberts, C. (1996) Information structure in discourse: towards an integrated formal theory of pragmatics. In *Working Papers in Linguistics*, pp. 91–136, Ohio State University Department of Linguistics
23. Levy, R. (2008) A noisy-channel model of rational human sentence comprehension under uncertain input. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 234–243, Association for Computational Linguistics
24. Bergen, L. and Goodman, N.D. (2015) The strategic use of noise in pragmatic reasoning. *Top. Cogn. Sci.* 7, 336–350
25. Franke, M. and Jäger, G. (2016) Probabilistic pragmatics, or why Bayes rule is probably important for pragmatics. *Z. Sprachwissenschaft* 35, 3–44
26. Lewis, D. (1969) *Convention: A Philosophical Study,* John Wiley & Sons
27. Frank, M. and Goodman, N. (2012) Predicting pragmatic reasoning in language games. *Science* 336, 998
28. Qing, C. and Franke, M. (2015) Variations on a bayesian theme: comparing bayesian models of referential reasoning. In *Bayesian Natural Language Semantics and Pragmatics*, pp. 201–220, Springer
29. Carstensen, A. *et al.* (2014) Testing a rational account of pragmatic reasoning: the case of spatial language. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, pp. 2009–2013, Cognitive Science Society
30. Degen, J. and Franke, M. (2012) Optimal reasoning about referential expressions. In *Proceedings of SemDIAL*, La Region Île de France/CLILLAC-ARP/Laboratoire Linguistique Formelle/LabEx Empirical Foundations of Linguistics (EFL)
31. Vogel, A. *et al.* (2013) Implicatures and nested beliefs in approximate decentralized-pomdps. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pp. 74–80, Association for Computational Linguistics
32. Franke, M. and Degen, J. (2016) Reasoning in reference games: individual vs. population-level probabilistic modeling. *PLoS One* 11, e0154854
33. Camerer, C.F. *et al.* (2004) A cognitive hierarchy model of games. *Q. J. Econ.* 861–898
34. Bergen, L. *et al.* (2012) That's what she (could have) said: how alternative utterances affect language use. In *Proceedings of the 34th Annual Conference of the Cognitive Science*, Cognitive Science Society
35. Degen, J. *et al.* (2013) Cost-based pragmatic inference about referential expressions. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, pp. 376–381, Cognitive Science Society
36. Nordmeyer, A.E. and Frank, M.C. (2014) A pragmatic account of the processing of negative sentences. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, Cognitive Science Society
37. Graf, C. *et al.* (2016) Animal, dog, or dalmatian? Level of abstraction in nominal referring expressions. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, Cognitive Science Society
38. Goodman, N.D. and Stuhlmüller, A. (2013) Knowledge and implicature: modeling language understanding as social cognition. *Top. Cogni. Sci.* 5, 173–184
39. Breheny, R. *et al.* (2013) Taking the epistemic step: toward a model of on-line access to conversational implicatures. *Cognition* 126, 423–440
40. Kao, J.T. *et al.* (2014) Nonliteral understanding of number words. *Proc. Natl. Acad. Sci. U. S. A.* 111, 12002–12007
41. Kao, J.T. and Goodman, N.D. (2015) Let's talk (ironically) about the weather: modeling verbal irony. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, Cognitive Science Society
42. Kao, J.T. *et al.* (2014) Formalizing the pragmatics of metaphor understanding. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, pp. 719–724, Cognitive Science Society
43. Williamson, T. (2002) *Vagueness,* Routledge
44. Lassiter, D. and Goodman, N.D. (2015) Adjectival vagueness in a bayesian model of interpretation. *Synthese* 1–36
45. Geurts, B. and Pouscoulous, N. (2009) Embedded implicatures?!? *Semant. Pragmat.* 2, 4–1
46. Chemla, E. and Spector, B. (2011) Experimental evidence for embedded scalar implicatures. *J. Semant.* 28, 359–400
47. Potts, C. *et al.* (2015) Embedded implicatures as pragmatic inferences under compositional lexical uncertainty. *J. Semant.* Published online December 18, 2015. http://dx.doi.org/10.1093/jos/ffv012
48. Bergen, L. *et al.* (2016) Pragmatic reasoning through semantic inference. *Semant. Pragmat.* 9, 1–83
49. Bar-Hillel, Y. (1971) Out of the pragmatic wastebasket. *Linguist. Inq.* 2, 401–407
50. Jones, M. and Love, B.C. (2011) Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behav. Brain Sci.* 34, 169–188
51. Fox, D. and Katzir, R. (2011) On the characterization of alternatives. *Nat. Lang. Semant.* 19, 87–107
52. Monroe, W. and Potts, C. (2015) Learning in the rational speech acts model. In *Amsterdam Colloquium*, ILLC
53. Marr, D. (1982) *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information,* Henry Holt and Co
54. Degen, J. and Tanenhaus, M.K. (2015) Processing scalar implicature: a constraint-based approach. *Cogn. Sci.* 39, 667–710
55. Grodner, D.J. *et al.* (2010) Some, and possibly all, scalar inferences are not delayed: evidence for immediate pragmatic enrichment. *Cognition* 116, 42–55

56. Golland, D. *et al.* (2010) A game-theoretic approach to generating spatial descriptions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 410–419, Association for Computational Linguistics
57. Vogel, A. *et al.* (2013) Emergence of Gricean maxims from multi-agent decision theory. In *Human Language Technologies: The 2013 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 1072–1081, Association for Computational Linguistics
58. Andreas, J. and Klein, D. (2016) Reasoning about pragmatics with neural listeners and speakers. In *EMNLP 2016*.
59. Orita, N. *et al.* (2014) Quantifying the role of discourse topicality in speakers' choices of referring expressions. In *ACL 2014*, p. 63, Association for Computational Linguistics
60. Vouloumanos, A. *et al.* (2012) Twelve-month-old infants recognize that speech can communicate unobservable intentions. *Proc. Natl. Acad. Sci. U. S. A.* 109, 12933–12937
61. Clark, E.V. and Amaral, P.M. (2010) Children build on pragmatic information in language acquisition. *Lang. Linguist. Compass* 4, 445–457
62. Noveck, I. (2001) When children are more logical than adults: experimental investigations of scalar implicature. *Cognition* 78, 165–188
63. Papafragou, A. and Musolino, J. (2003) Scalar implicatures: experiments at the semantics-pragmatics interface. *Cognition* 86, 253–282
64. Barner, D. *et al.* (2011) Accessing the unsaid: the role of scalar alternatives in children's pragmatic inference. *Cognition* 118, 84–93
65. Frank, M.C. and Goodman, N.D. (2014) Inferring word meanings by assuming that speakers are informative. *Cogn. Psychol.* 75, 80–96
66. Horowitz, A.C. and Frank, M.C. (2016) Children's pragmatic inferences as a route for learning about the world. *Child Dev.* 87, 807–819
67. Keysar, B. *et al.* (2003) Limits on theory of mind use in adults. *Cognition* 89, 25–41
68. Lane, L.W. *et al.* (2006) Don't talk about pink elephants& speakers' control over leaking private information during language production. *Psychol. Sci.* 17, 273–277
69. Gatt, A. *et al.* (2013) Are we bayesian referring expression generators. *Proceedings CogSci* (Vol. 35), Cognitive Science Society
70. Baumann, P. *et al.* (2014) Overspecification and the cost of pragmatic reasoning about referring expressions. In *Proceedings of the Thirty-Sixth Annual Conference of the Cognitive Science Society*, Cognitive Science Society
71. Levelt, W.J. (1993) *Speaking: From Intention to Articulation,* MIT Press
72. Yoon, E.J. *et al.* (2016) Talking with tact: polite language as a balance between kindness and informativity. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, Cognitive Science Society
73. Smith, N.J. *et al.* (2013) Learning and using language via recursive pragmatic reasoning about other agents. In *Advances in Neural Information Processing Systems*, Curran Associates
74. Piantadosi, S.T. *et al.* (2012) The communicative function of ambiguity in language. *Cognition* 122, 280–291
75. Kirby, S. *et al.* (2008) Cumulative cultural evolution in the laboratory: an experimental approach to the origins of structure in human language. *Proc. Natl. Acad. Sci. U. S. A.* 105, 10681–10686
76. Perfors, A. and Navarro, D.J. (2014) Language evolution can be shaped by the structure of the world. *Cogn. Sci.* 38, 775–793
77. Kirby, S. *et al.* (2015) Compression and communication in the cultural evolution of linguistic structure. *Cognition* 141, 87–102
78. Regier, T. *et al.* (2007) Color naming reflects optimal partitions of color space. *Proc. Natl. Acad. Sci. U. S. A.* 104, 1436–1441
79. Kemp, C. and Regier, T. (2012) Kinship categories across languages reflect general communicative principles. *Science* 336, 1049–1054
80. Xu, Y. and Regier, T. (2014) Numeral systems across languages support efficient communication: From approximate numerosity to recursion. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, Cognitive Science Society