Understanding the models

1. What is the status of RSA as a framework? Given the diversity among models, what are the core features of RSA?

The RSA modeling framework views language understanding as recursive social reasoning between speakers and listeners. Language users interpret utterances by reasoning about the process that generated that utterance: a model speaker trying to inform a naive listener about the state of the world. Language users choose utterances in a similar fashion, by reasoning about how a listener would interpret the utterance they choose. In practice, these various levels of reasoning get implemented as simulation-based probabilistic programs, here in the probabilistic programming language WebPPL.

We purposefully describe RSA as a framework rather than merely a (family of) model(s) in order to highlight its status as a mode of inquiry. By formally articulating the pragmatic reasoning process, RSA serves as a tool for understanding how language works. RSA allows us to move beyond high-level prose descriptions of pragmatic phenomena to actually *implementing* the reasoning via simulation-based probabilistic programs. When successful, these models serve as a proof-of-concept for the computational-level description of the relevant phenomena. But they become even more interesting when they fail. Because the models make clear qualitative and quantitative predictions about language use, they can be tested against behavioral data. Where the patterns predicted by the model do not match those observed empirically, model assumptions must be incrementally revised. These revisions serve as testable hypotheses concerning the ways in which various aspects of context interact in communication. What results is a seemingly-diverse family of models covering a range of different pragmatic phenomena.

Despite this diversity, the models we have encountered share various features; it is these features that stand at the core of the RSA framework. The grounding assumption of RSA is that language gets used for the purpose of communication. The goal of the speaker is to successfully transmit information to a listener; the goal of the listener is to successfully recover that information from the speaker's utterance. The details of this process -- nested inference about utterances, integrating world knowledge in context -- define the RSA framework.

The pragmatic calculus proceeds in a similar fashion for each of the models we encountered: a speaker observes some state and chooses an utterance to best communicate that state to a listener in order to resolve some Question Under Discussion (QUD); the listener observes that utterance and infers the state that the speaker must have encountered. Thus, in the RSA framework, the speaker and listener coordinate on the utterance and interpretation that are most likely to correctly resolve the QUD.

In order to formally articulate the pragmatic reasoning that stands at the heart of RSA, our models minimally require the following components. First, we must model the various levels of inference. A typical RSA model includes a literal listener, which serves to ground the truth-functional semantics; a speaker, who chooses utterances by soft-max optimizing utility; and the pragmatic listener, who integrates their beliefs about the speaker and the world via Bayes' rule to arrive at an interpretation of the observed utterance. Second, we must model the prior beliefs that language users bring to bear on conversation itself: the possible states and their relative probabilities, the possible utterances and their relative costs, and a truth-functional mapping between the utterances and the states they describe. Third, we must model additional aspects of the utterance context, including the relevant QUD(s) and beliefs about speaker optimality. Any number of these model components may be modified in various ways, but the core architectural assumption of nested inference about utterances integrating world knowledge in context remains.

2. How does one decide which aspects of context to represent in a model?

The question of what counts as context is an unsolved -- and indeed unsolvable -- problem. There is an infinity of facts about any communicative scenario: the color of the floor, the temperature of the room, the number of the

attendant cats, etc. Which facts matter for the purpose of communication? The goal of modeling interactive language use cannot be to represent the entirety of the world -- the world does that already for us. Instead, the goal is to identify, isolate, and represent the interaction of those aspects of context that influence communication in systematic ways.

RSA does not answer the question of which aspects of context influence communication. However, it does offer a way to explicitly represent the interaction of those aspects of context that the modeler hypothesizes matter in the computation of meaning, all while assuming a central mechanism by which meaning is computed: Bayes' rule. The resulting model generates quantitative predictions which can then be tested against behavioral data. Via subsequent model comparison, RSA thus provides the means by which to ask the question of which aspects of context influence communication.

Across the models that we have encountered, the aspects of context that are represented by default are the relevant state space, prior beliefs about the state space (world knowledge), as well as utterance alternatives and their properties. RSA models also contain implicit assumptions about the QUD, properties of the speaker, and the contents of common ground, even without explicit alternatives. For instance, unless modeled otherwise

- the representation of the state space implicitly encodes the QUD that an utterance's informativity is computed with respect to,
- the speaker is represented as a cooperative speaker with the goal to soft-maximize utility (by computing utterance informativity),
- world knowledge is implicitly assumed to be in common ground by having the speaker and the listener perform their computations on the same prior, and
- utterance alternatives are implicitly assumed to have equal costs.

But RSA also allows for explicitly modeling these factors. In Chapters 3 and 4, we saw how introducing explicit QUD alternatives changes the result of the informativity computation, but not the general formalization of the computation of that informativity. Similarly, in Chapters 2 and 6, we saw how explicitly modeling a non-omniscient speaker increases the listener's uncertainty, but not the computations performed to resolve that uncertainty. In Chapter 8, we saw that the amending the speaker's utility function can accommodate social goals beyond informativity. Degen, Tessler & Goodman 2015 show how deviating from the assumption that priors are always in common ground can capture the situation-specific effects of prior knowledge. In Chapters 4 and 6 , we considered a principled way by which to assign asymmetric costs to utterances. And in Chapter 7, we saw how to model richer world knowledge via a structured prior.

It is an empirical question whether representing these aspects of context will suffice for the purpose of characterizing pragmatic language use. Probably not. Still, the above factors are likely to matter in the computation of any meaning-related inferences. And, the true power of RSA is not in isolating these factors, but in offering a principled way in which to formulate their interaction via nested probabilistic inference. Forcing researchers to be explicit about the precise way in which these factors interact allows for the systematic investigation of contextual information in communication.

3. We know that language users are limited cognitive agents. Where do resource limitations come in?

In short, the answer to this question is: across models, resource limitations are captured in the rationality parameter alpha and in the assumed depth of reasoning. In both cases, these aspects of the models are mere approximations of the resource limitations; we are not modeling the limitations themselves, but rather the downstream effects they have on the reasoning process. Resource limitations may also be captured by model-specific assumptions, for example,

the prior on quantifier scope we saw in Chapter 4, which we interpreted as differences in the ease with which a particular scope assignment comes to mind a priori.

If you are someone interested in memory, attention, cognitive control, and related issues, you will find this a very unsatisfying answer. In general, there is a distinct lack of explicit engagement of RSA modelers with issues of resource limitations. This is a result of RSA models being an approach in the tradition of rational analysis (Anderson, 1991), an approach towards formulating and revising theories of cognition. Anderson specifies rational analysis as proceeding in the following six steps:

1. Specify precisely the goals of the cognitive system.
2. Develop a formal model of the environment to which the system is adapted.
3. Make the minimal assumptions about computational limitations.
4. Derive the optimal behavioral function given 1-3 above.
5. Examine the empirical literature to see whether the predictions of the behavioral function are confirmed.
6. Repeat, iteratively refining the theory.

The focus in the rational analysis approach is clearly on what the agent's optimal behavior should look like. However, the fact that rationality is bounded is acknowledged in step 3. The minimal assumptions about computational limitations included in any RSA model are non-infinite values of the rationality parameter alpha and non-infinite depths of recursion. The smaller the alpha parameter, the less the speaker maximizes their utility on a particular reasoning step; and the lower the depth of recursion, the less overall reasoning there is. While alpha values generally vary wildly across tasks -- indeed, one should never interpret a alpha value in isolation -- most RSA models assume the listener is an L1 listener and the speaker an S1 speaker.

Thus far, little attention has been paid to an exploration of the tradeoff between depth of recursion and alpha value (but see Qing & Franke 2015, and Franke & Degen 2016). In some cases, a higher alpha value results in the same patterns as a deeper level of recursion. For example, in scalar implicature a higher alpha value leads to more extreme probabilities on the speaker and consequently on the listener side. The same qualitative effect (with quantitative differences) is achieved by assuming a lower alpha value but a greater recursive depth. In other cases, where the space of alternatives is larger, or the resulting belief distributions less peaked, etc., the effect of adding a layer of recursion may be different. This is an issue that requires further systematic exploration.

In general, we believe that connecting the black box alpha to the psychological literature on memory, attention, and cognitive control is a worthy enterprise. It may also turn out that resource limitations must be modeled in other ways; but thus far, capturing them in the above terms appears to be sufficient. It also has the advantage of not having to assume any particular mechanism for resource allocation, allowing for a more direct focus on the pragmatic reasoning process itself.

Here, another question that commonly arises is whether the assumed reasoning process itself is too sophisticated.. The claims that are typically made is that language users aren't capable of deep recursive reasoning, or that it is much too complicated to integrate so many different pieces of information in real time. If one were to interpret these models as models of the online reasoning process, then it would be an empirical question whether this reasoning process is too complicated to be executed online during conversation. However, since its original conception, RSA modelers have been careful to point out that the models are intended to provide a computational level analysis of the language use problem that agents face (Marr, 1982), rather than specifying the mechanism by which they solve that problem. One should not take, e.g., each level of recursion as an actual reasoning step. If one wanted to interpret this models as online processing models, one would need to formulate linking hypotheses between model output and measures of online processing. One should avoid making pronouncements about presumed processing difficulty simply on the basis of architectural considerations of these models without such linking hypotheses. With these

hypotheses in hand, we can test the question of whether reifying any model component as a psychological mechanism is useful for explaining online processing phenomena.

Testing the models

- How do we test the predictions of these models?
  - When do we need to empirically measure the prior (vs. assume a uniform prior)? What are different ways of estimating the prior?
  - Is there a way of measuring what the QUD is? Or, manipulating the QUD..
  - Let's say we want to do philosophy of science on RSA: to what extent are we worried that you can make RSA "do anything"? (RSA as a methodological tool vs RSA as a theory of language use and interpretation)
- Bayesian inference is intractable in the limit

Extending the models

- How do we extend RSA to language learning?

- How do we add syntax or more rich language structure?
  - How should one do compositionality (especially for soft semantics)?

- Can we think about cross-cultural/linguistic differences in terms of different weights on different components of the model?

- What are the challenges that we expect in integrating time -- both by going sub-sentential and model incremental belief update and by going supra-sentential and model dialog, and even larger to linguistic conventions?

- XXX unified cost function? XXX interlocutor-specificity/adaptation?