Towards formal theories and computational models of evolving referents to unfamiliar targets

Veronica Boyce

March 5, 2024

Language is an amazing human technology that supports much of human culture by being an efficient way of conveying many concepts with precision. Humans communicate in a range of situations, using language in ways that range from scripted and stereotyped (ex. greetings) to very open ended. The extensibility of language for use in novel contexts supports the transmission of new ideas and communication with new people.

Much linguistic creativity is ephemeral, occurring at at a single time within a single conversation to fulfill a one-time communicative need. However, some linguistic needs are recurrent, and the words used to fill these needs may eventually conventionalize and lexicalize within the community. If the communicative niche occurs across groups, the conventions can spread to larger communities. At the long time scale, the stickiest linguistic innovations make up language evolution.

Whether descriptions are fleeting or lasting, the core observation is the same: humans are adept at using language to communicate about unfamiliar targets with each other. This success entails two skills: a) the ability to refer to a new thing (that doesn't have an established name) in a way an interlocuter can understand and b) the ability to refer to a not-quite-so-new anymore thing in a way that is sensitive to the conversation history and may eventually converge into a conventional name.

How can we explain and predict these phenomena? Following Meehl (1990), we will prefer theories that make more specific (and thus more testable and risky) predictions. In particular, models that make predictions about functional forms make stronger connections to theories are other levels of analysis. We will prefer computational and mathematical models because these more clearly make predictions and are explicit about their degrees of flexibility (Yarkoni & Westfall, 2017). [Veronica promises a more fleshed out philosophy of science interlude here in the actual dissertation, but needs to read more to have better backed up takes first.]

1 Levels of explanation

One issue with developing formal models of human linguistic behavior is that the data are rich and the possibilities are vast. How can one take rich, messy interaction data and tame it into meaningful features that are tractable to model and predict?

To ground this, the common experimental paradigm for exploring communication about non-codeable items is some type of iterated reference game. One person (the director) has privileged knowledge about a target item or target order and must communicate this to one or more listeners so they can identify the target or reorder their images to match. The same images are used repeatedly for multiple rounds of matching. The primary questions of interest are how people use language to succeed at this task and how their language use changes over the course of the interaction, as well as how language use changes under different variants (ex. switching out listeners / speakers / items).

One can define the question of interest, and thus the scope of the modeling and prediction problem, at a few different levels:

1. A clear trend observed in the iterated reference games is that utterance length reduces over repeated reference to the same novel target between the same pair of interlocuters. This has a straightforward output metric (utterance length in words), and given enough data, a functional form could be fit. Determining the relevant input predictors, and operationalizing them would be a challenge, and then there's the wide variation in reduction rate for individual dyads within an experimental condition. This level could be the target of statistical models, but may not be well-suited to explanatory theories.

2. Instead of predicting the entire curve of a conversation, we could instead ask: given the environment and the conversation history thus far (i.e. any previous references to the target), can we predict properties of the next description? While still a high-level model, this at least parallels the presumed generative process of people selecting utterances and actions based on what has occurred thus far. Group-to-group variation in the dynamics may be smaller and only integrate into larger differences over time. This approach can also be neatly separated into questions of how to model initial utterances and separately the dynamics of how successful utterances evolve over repetition.

3. Finally, one may wish to model the actual production and comprehension processes. We might wish to predict the incremental understanding of a new reference expression given conversation history and context. Conversely, we might want to model the unit by unit production of a referring expression. This level of analysis could be seen as the algorithmic counterpart to the computational level approach in 2. Notably addressing these fine-grained process questions would require correspondingly detailed time course data to predict.

Currently, we have formal models at none of these levels. Personally, I find the 2nd level the most promising, but ideally, we would have mutually compatible formal theories at all levels. In the rest of this introduction, I will discuss the current state of the literature and gesture at what types of work will be needed to bridge towards formal models.

At whatever level, we would like to eventually have theories that are precise in what they predict (e.g. preferring quantitative predictions to directional ones) and clear in how generalizable they are (e.g. is the theory only about face to face communication between dyads, or does it make claims over a wider range of experimental settings).

I will cover approaches from four perspectives. First, the communication and reference game literature which provides descriptive characterizations of phenomena of interest. Next, I will discuss two optimization-oriented frameworks of efficiency and the Rational Speech Acts models. Each of these might be able to supply formal theory, although as we will see, there are significant challenges in aligning these frameworks with the rich, open-ended language data. Finally, I will touch on the psycholinguistic approach, reveals some constraints that computational/optimization approaches may need to account for.

2 Communication

The communication and conversation literature, especially with regard to referring expressions, has provided useful descriptive characterizations of how people communicate in naturalistic and experimental settings. There is an accumulation of fairly consistent and reliable findings of reduction, convention, and partner-specificity that could be the targets of theorizing. These describe consistently observable trends, but they are difficult to compare quantitatively across studies because of lack of precise specification and mathematical models.

Our understand and predictive power around these phenomena is limited. We do not know how they respond to "twiddling the experimental knobs", nor are there theories that make risky, testable predictions about necessary and sufficient conditions or the functional form. We can describe the trends of what happens in the range of conditions that have been experimented on, but theories rarely make predictions of what would happen in other conditions or what experiments would adjudicate between theories.

2.1 Mentalizing versus Non-mentalizing approaches

A big question that comes up with conversation, and interactions between agents more generally, is whether and how agents track other agents internal states of knowledge and how this factors into their interaction.

The "mentalizing" tradition treats humans as representing other humans as agents with internal states that include knowledge and goals. Within this broad school, there is variation in how these representations are implemented, how information gets added or modified, what exactly is tracked, and when representations (versus heuristics) are used.

Within this tradition, many use the term "common ground" to refer to knowledge that two agents share. In some cases, it is used in a pre-theoretic way to mean roughly "things you think another person will understand and won't be surprised if you reference" (Garrison et al., 2022; Leung et al., 2023). For instance, Hanna et al. (2003) defines common ground as the "mutual knowledge, beliefs, and assumptions" held by the interlocuters. This meaning is roughly comparably to "givenness" in other domains (Fay et al., 2010).

However, the problem with the term "common ground" is that some use it in a theoretically very loaded way, originating from the privileged versus mutual versus common knowledge framework (Clark, 1996). Under this usage, "common ground" is defined via infinite recursion in knowing that the other person knows that the first person knows that ...; this is the usage that comes up in formal semantics where many things may be introduced to common ground via accommodation (Horton & Keysar, 1996; Pickering & Garrod, 2004). In practice, humans don't tend to do more than a couple layers of recursion in their pragmatic reasoning (Franke & Degen, 2016). Thus, it is generally not important to distinguish knowledge types at deeper recursion levels than mutual knowledge that both people know to be mutual.

How do we determine that something is mutually known with another person? Many approaches have tried to characterize when something is mutually known (Brown-Schmidt, 2012; Clark, 1996; Horton & Keysar, 1996). This has predominately taken a deterministic approach with rules such as "if it's in shared visual presence, it's mutually known". Enumerating all the options doesn't work because we don't always have certainty around what is or isn't mutually known; someone can look at something in shared visual presence, but not know about it because they weren't attending. Visual presence may be a good heuristic, but good heuristics won't be perfect, and humans operate under uncertainty. Understanding human communicative behavior also doesn't require a deterministic answer to when something is mutually known: what another person knows or can be expected to understand is something that computational models will want as an input or intermediary, so that it can be used to evaluate utterance options. However, the knowledge state can clearly be probabilistic and may be inferred from empirical data.

The mentalizing approaches can be contrasted with "interactive alignment theory" which attempts to explain how people can successfully collaborate on reference tasks without reasoning about each other's mental states (Gandolfi et al., 2022; Pickering & Garrod, 2004). The motivation for non-mentalizing accounts is the apparent difficulty of mentalizing, coming from accounts such as naive egocentrism, naive realism, and initial egocentrism [Veronica promises citations for this, but later]. These intellectual traditions claim that mentalizing is a mentally taxing add-on, that is computationally expensive and not automatic. Under this framework, (Gandolfi et al., 2022; Pickering & Garrod, 2004) try to account for observed alignment between interlocuters by extending ideas such as lexical alignment [Veronica again promises citations].(Pickering & Garrod, 2004) claims that the alignment occurs via "priming" and is "resource-free and automatic", without providing a further explanation of what this means or how this works in terms of memory and processing. Given that humans reason socially about each other readily and from a young age (Rakoczy, 2022), it's not clear how well the motivation for a non-mentalizing approach holds up. Additionally, the interactive alignment account seems unable to explain reduction phenomena where the expressions change.

As an aside, the "common ground" tradition and the "interactive alignment" traditions have tended to use different types of experiments, with "common ground" generally using asymmetric director/matcher designs (dating back to at least Krauss and Weinheimer (1966)) and the "interactive alignment" traditions using symmetric designs such as the 'maze' task. Thus it is possible the two approaches are build around trying to explain differing sets of experimental results.

2.2 Partner specificity, audience design, and sharing effort

One key phenomenon from iterated reference games to unfamiliar images is that switching matchers or adding a new matcher changes the describers behavior, as they shift to longer descriptions. This change in behavior is described as "partner-specificity", with the idea being that the conventional names developed with one partner as specific to that partnership (Brennan & Clark, 1996; R. D. Hawkins, Liu, et al., 2021; Metzing & Brennan, 2003). The idea of partner-specificity is also referenced with regard to how different pairs diverge to different names for the targets (R. D. Hawkins et al., 2020). Partner specificity is part of the mentalizing tradition and assumes that partners (and their background and knowledge states) are being represented in the minds of speakers. The form of the representation is not explicitly stated, so these models could be compatible with heuristic or distributed representations, but include explicit thinking about the audience and are incompatible with the non-mentalizing "priming" account.

Empirical evidence from experiments where one director talks with multiple partners suggests that people do "partial pooling" over their partners (R. D. Hawkins, Franke, et al., 2021; S. O. Yoon & Brown-Schmidt, 2014). That is, a speaker A will show some variation in their expressions when talking to partner B versus partner C, but there will be some generalization between partners as well, so that A talking with B is more like A talking with C than D talking to E. When coupled with a tendency for descriptions to shorten within a pair, this leads to a jagged pattern of reference length: when switching to a new partner, speakers use longer utterances, but not as long as their initial utterance with their first partner (S. O. Yoon & Brown-Schmidt, 2019b).

A related term is "audience design", the idea that speakers seem to be sensitive to the knowledge state of their listener and say things that are easy for the listeners to comprehend. Confusingly, "audience design" sometimes implies intention on the part of the speaker (Horton & Gerrig, 2002, 2005), and sometimes is used when utterances are constructed based on what's easy for the speaker, and listener ease is a side effect (Horton & Keysar, 1996; MacDonald, 2013; Rogers et al., 2013). For instance, audience design could both occur in times when the speaker gives an elaborated description to a naive listener (inferred to be intentional if contrastive with their description to a non-naive listener), but speakers may also tend to start descriptions with given material which is both more accessible to the speaker and convenient for the listener. Intention versus side-effect are difficult to distinguish between because speakers and listeners often share recent context, find the same things salient, and linguistically what is easier to produce is often easier to process. Thus, disentangling speaker and listener ease may require careful experimental designs where ease of production and ease of comprehension are separated (Ferreira, 2004).

Questions around audience design are related to larger issues of how interlocuters split the communicative burden with one another. Depending on the task and the communication modality, there may be many options for how to balance the communicative load (Clark, 1996; Fay et al., 2010; Fox Tree & Clark, 2013). For instance, a listener could describe what options they see or otherwise prompt the speaker. We might expect the load splitting to vary based on the capacities of the interlocuters (ex. a speaker might craft their utterances more when talking to a child versus an adult) and the capacities of the channels (ex. speakers may use different approaches if listeners can interrupt).

Multi-way conversations complicate the verbal theories of audience design and partner specificity by introducing a larger audience of more partners. Two main questions are whether "aim low" or "aim high" in balancing the needs of the listeners and whether speakers track individual listeners or an aggregate (S. O. Yoon & Brown-Schmidt, 2014). Empirical results indicate that speakers are sensitive to the knowledge states of listeners in a gradient way (S. O. Yoon & Brown-Schmidt, 2014, 2018; S. O. Yoon & Brown-Schmidt, 2019a). At least in small groups, speakers can track the correspondence between individual listener identity to histories and knowledge states, and can incorporate contextual factors that modulate task difficulty into their considerations (S. O. Yoon & Brown-Schmidt, 2019b). Speakers also take strategies in group contexts that don't occur as often in dyadic contexts, such as referring to a target with both the name that one person will understand and a elaborated description that will help another person get on the same page (S. O. Yoon & Brown-Schmidt, 2018). The ability to track partner's knowledge states presumably would degrade as groups got bigger, but this paradigm has not been used for groups large enough for this to happen.

2.3 Convention formation

Over repetitions with the same partner, dyads in repeated reference games tend to form shared "conventions" (also called "conversational pacts") about how to refer to the initially ambiguous targets. These conventions tend to be partner- and context-specific: changes in the speaker, audience members, or changes in the context can all license the use of a new description (Ibarra & Tanenhaus, 2016; Metzing & Brennan, 2003; S. O. Yoon & Brown-Schmidt, 2014).

What exactly does convention formation refer to? There is ambiguity about what level of specificity convention formation and conceptual pacts refer to. It could be on the lexical level, such as calling a figure “ballerina”. It could be conceptualizing the figure as a ballet dancer with a tutu (manifesting in descriptions with semantic association, but not lexical overlap, such as “ballerina” and “dancing in a tutu”). It could also be a general paradigm for how to describe figures, such as in terms as humans in different postures. Horton and Gerrig (2002) distinguish between “lexical entrainment” when the same words are reused, and “conceptual similarity” when there is broader similarity that does not repeat the same words. These levels often co-occur, but in order to have a computational model of the phenomenon, we need to be clear about which is meant in order to operationalize it.

The semantic meaning of a description is not a priori related to its length, but these two features empirically tend to correlate in iterated reference games (R. D. Hawkins et al., 2020). Thus “reduction” or the shortening of utterances is sometimes used as a shorthand and measurement proxy for the semantic changes (Clark & Wilkes-Gibbs, 1986; R. D. Hawkins, Franke, et al., 2021). Convention and reduction are sometimes conflated with partner-specificity, as these phenomena often co-occur with different pairs forming different conventions and changes in group composition leading to (temporarily) longer descriptions (Clark & Wilkes-Gibbs, 1986; Wilkes-Gibbs & Clark, 1992).

It remains an empirical question whether the shortening of utterances, convention formation, and partner-specificity of descriptions are inseparable or merely occur together in the experimental paradigms considered in the literature.

Testing these patterns would require studying the phenomena in more varied experimental situations. One question is how convention formation might change with group size, as some of the pieces, like partner-specificity seem like they will have to fall apart if groups grow to dozens of people. One angle on larger groups is network structures, where (non-linguistic) convention formation has been studied on networks of up to 50 people (Guilbeault et al., 2021). In this communication game with targets varying over continuous space, large groups tend to end up with fairly consistent category boundaries across independent networks, while small networks can support more idiosyncratic category boundaries. This work is suggestive that there may be group-size and group/network-structure dependencies on what sort of conventions arise. There has been some work on group and network designs in traditional iterated reference games, but only on relatively small groups.

Typically, in convention formation, holistic or analogic descriptions are the ones that stick (Clark & Wilkes-Gibbs, 1986), but this isn’t absolute. Groups can successfully coordinate on reference using many different types of names, including ones that pick up on low-level or meta-level features. The range of successful options makes explaining convention formation harder, and suggests suggests a strong path dependency for how reduced utterances evolve, potentially influenced by factors such as relationships and humor value.

2.4 Developmental work

Many of the phenomena discussed above are likely to depend on varied communicative and linguistic skills. One question therefore is at what age these phenomena emerge, as that provides insight into when the underlying skills must exist by. There has been relatively limited work on children, and much of it is hampered by small sample sizes and limited numbers of experiments. Thus, unlike the robustness of the key phenomena in the adult literature, there is uncertainty about the generalizability of the findings with children.

Children show rudiments of the key phenomena from an early age, although they remain unadultlike for a while. Children ages 3 to 5 show sensitivity to referential pacts, in protesting the breaking of a pact, although they sometimes protest even when a new speaker uses a new term (Matthews et al., 2010), in contrast to adults who allow new speakers to form new pacts (Metzing & Brennan, 2003). In a test of both context and partner-specificity (modeled on Brennan and Clark (1996)), 6 year olds showed the adult-like pattern of maintaining the same terms with a partner even when a context change rendered them overinformative, while 4 year olds struggled more on the task and did not show partner-specificity (Köymen et al., 2014). In addition to partner-specificity, adults are also sensitive to the context and whether distractor items are close or far. Evidence for children is mixed, with Abbot-Smith et al. (2016) finding some signs of sensitivity to context in 2 and a half year olds, but in a different paradigm 4-8 year old children did not appropriately modulate labels to familiar objects based on context (Leung et al., 2023).

In repeated reference games to novel objects, there are claims of young children's inability to form referential pacts, followed by a gradual increase in competence throughout childhood and adolescence, although these studies have methodological concerns that could mask earlier competence (Glucksberg et al., 1966; Glucksberg & Krauss, 1967). More recently, 8-10 year old children in Branigan et al. (2016) seem to demonstrate an ability to describe abstract shapes and some level of convention formation (measured via reduction), although children show mixed results in terms of sensitivity to an existing versus naive participant (compare adult performance in Wilkes-Gibbs and Clark (1992)).

Overall, the available literature indicates a gradual path towards adult-like behavior, but the limited number of studies means all these claims should be interpreted with caution. Understanding the developmental trajectory of pragmatic and communicative skills requires more rigorous and systematic data on children.

2.5 Takeaways

The communication and reference game literature provides descriptive theories that identify some phenomena of interest and raises questions around whether these phenomena occur intentionally or emerge as a by-product of other processes. Within the narrow range of experimental paradigms, clear patterns around reduction, convention, and partner-specificity are robustly observed, and could form a list of results that models could hope to explain and predict. Our current understanding of these phenomena is limited. We do not have ways to predict them quantitatively, nor do we have a grasp on the necessary or sufficient experimental conditions under which they occur (to what extent). In many cases, we do not even have consensus on how exactly the characterize the phenomena. In order to build strong theories, we first need clarity around what exactly the theories should be accounting for. A core theoretical question is whether the observed patterns of reduction, convention, and partner-specificity require "special" mechanisms, or whether these results can sufficiently be explained by broader coverage theories of efficiency and rational communication.

I next describe these broader coverage theories before returning to the question of how to join them to the phenomena of interest.

3 Efficiency

One unifying framework gaining traction in psycholinguistics is efficiency, the idea that language and language use is under pressure to support efficient communication by maximizing the ratio of relevant information transmitted to effort. Efficiency is a high level framework that requires a number of linking assumptions to render it testable against data; however, comparisons of attested language use to counterfactual options can bound what assumptions are needed for parsimony.

Efficiency is thought to arise from trade-offs between communicative expressivity and some combination of learnability and easy of production (Kirby et al., 2015; Piantadosi et al., 2012).

Evidence for efficiency comes from the argument that features of language are distributed much more closely to the Pareto frontier than would be expected by chance. A historically well-known example is that word frequencies follow a power-law distribution, which Zipf (1949) explains in terms of a "principle of least effort", although note that power-law distributions are common across domains and generated by a variety of processes (Piantadosi, 2014). Stronger evidence comes from the lexical partitioning of subdomains such as color, number, and kinship terms, where the distribution of systems falls on the frontier between complexity (number of terms) and informativity (how many bits each term provides) (Gibson et al., 2019; Kemp et al., 2018; Zaslavsky et al., 2018). Syntactic features of language such as harmonic word order and dependency length also appear to be optimized for increased expressivity with minimized processing effort (Gibson et al., 2019; J. Hawkins, 1995).

Efficiency arguments are based on the language artifacts of grammars and transcripts, but efficiency pressures act on language use as a process, not language as a static code (Gibson et al., 2019). Thus efficiency can be seen as imposing a joint constraint on the entire communicative process: minimizing the total time and effort involved in going from an idea in one person's head to a sufficiently close idea in another person's head. A corollary of this framing is that shorter utterances (as measured in syllables or clock-time) are not always efficient if they take longer to produce or parse.

3.1 Redundancy and over-informative referring expressions

How do we reconcile the evidence that language is efficient with violations of efficiency in language use ("look at the yellow banana") and the ubiquity of ambiguous utterances?

Terms such as "redundant" and "over-informative" are commonly used to describe situations where people produce modifiers that do not restrict the extension of a noun phrase (e.g. "blue cup" when only one cup is salient) (Rubio-Fernandez et al., 2021). People do produce these non-restricting modifiers in reference tasks, especially for color, but this behavior seems to run counter to the idea of efficient language use. Is this a contradiction?

Formalizing claims of redundancy or over-informativity requires a definition of what would be minimally informative, which in turn depends on a commitment to a fully specified semantic-pragmatic system. For instance, if specificity implicatures are within the option space, are those calculated before or after informativeness is measured (Bergen et al., 2016)? One could sidestep the thorny theoretical by empirically measuring the information content of different utterances by how they shift the entropy of the distribution of inferred meanings (Degen et al., 2020), but this does not scale up well.

The flip side of "redundancy" is ambiguity: many, many utterances are ambiguous. In general, strong contextual factors render the ambiguity a non-issue (Piantadosi et al., 2012), but this means we can't judge language outside of the physical and social context it is used in. Determining what is efficient language use in context requires not just analyzing phrases and their alternatives, but also how long utterances take to generate and comprehend, which may be highly contingent on contextual factors and conversational history.

The idea that utterances should have "just enough" information has inspired a wealth of empirical research into what utterances people produce and what utterances people comprehend. By comparing these two halves of language use, we can determine how calibrated utterances are to what other people will understand.

3.2 Takeaways

Sometimes, judging whether something is efficient or not may be clear cut, if all reasonable sets of assumptions return the same result. It is perhaps easier to judge something to be inefficient if there is a shorter alternative that is both easier to produce and easier to comprehend. In the general case, however, efficiency is very hard to cache out in specific predictions because of the many time scales the pressures operate on. What's efficient for an utterance in isolation may not be efficient when considered over an entire life of language use. Thus, the efficiency framework is not directly testable, but it's goodness as a theory instead relies on the parsimony of the linking theories that are required to meld it to the data.

The formation of conversational pacts is often characterized as efficient, but this claim has not been cashed out in formal models like those described above (Clark & Wilkes-Gibbs, 1986; R. D. Hawkins et al., 2020). "Efficiency" is an informal description of these pacts that captures that shorter expressions are *more* efficient, but a given shorted expression could still be mis-calibrated to a particular situation. It could be too short, leading to misunderstanding, or still longer than it needs to be for effective communication. Mis-calibration of this type could contradict claims about partner specificity of reduction, but would not have to. Communicators could intend to be optimally calibrated but run into production difficulties, masking their competence. Either way, current claims of efficiency in conversation do not connect directly with the formal claims of optimal efficiency in the "language design" literature, and it is unknown whether speakers and listeners are calibrated or not.

4 Rational Speech Acts Models

Rational Speech Acts (RSA) is an information-theoretic, computational framework for making quantitative predictions about pragmatic inferences in context (Frank & Goodman, 2012; Goodman & Frank, 2016). In principle, the pragmatic communicative behavior of reference games is a key candidate for modelling with RSA frameworks, but degrees of flexibility and the issue of scaling to open classes of descriptions and the corresponding need for flexible semantics may prove challenges.

The basic idea of the RSA family of models is to picture two interlocuters recursively reasoning about how the other would produce or interpret utterances, grounding out in a listener (or speaker) who behaves in a pre-specified "literal" way.

Computational frameworks such as RSA provide a way to factor together different trade offs and determine their relative weights in a model (Goodman & Frank, 2016). A softmax is taken over the scores of the options to produce a distribution of interpretations and utterances, with some parameters such as the degree of optimality fit based on data. This framework is usually run with one or two levels of recursion, where it tends to produce a reasonable fit to human experimental judgments, consistent with work finding that most people reason pragmatically at a low recursion depth (Franke & Degen, 2016).

RSA models have been used to predict some instances of ad hoc pragmatics as well as conventionalized pragmatic implicatures (Bergen et al., 2016; Degen et al., 2020; Goodman & Stuhlmüller, 2013). Models generally include a utility or informativity term that relates to how well an utterance resolves uncertainty in favor of the target referent. It is common to also include factors such as the prior likelihood of referring to each target (salience prior) and some cost on utterances where longer or more complex utterances are penalized (Goodman & Frank, 2016). Some models also go beyond informativity, incorporating options to infer the question-under-discussion (Kao, 2014; Qing et al., 2016) or for speakers to balance informativity with politeness (E. J. Yoon et al., 2018).

A full RSA model would incorporate all of these components and infer their weights. However, for tractability, usually only those features that are considered relevant to the domain of interest are included. Because of the flexible framework, it is possible to model many sources and levels of uncertainty, and then integrate out that uncertainty to make predictions, but also update on the sources of uncertainty in response to input.

4.1 The Challenge of Semantics

Perhaps the largest challenge to RSA models is the question of how to ground out the models in a "literal" listener or speaker. For the most part, RSA is tested in toy domains where the set of possible utterances are small and it is possible to enumerate a set of meanings (Bergen et al., 2016; Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013). For instance, in some domains, a soft or continuous semantics is used to represent that some dimensions of meaning might be more strongly informative than others (Degen et al., 2020). This semantics supports the prediction of patterns of "redundant" color adjective use in referring expressions.

Soft semantics can run into conflict with compositionality: either every possible utterance must independently receive a degree of match with every possible object in the prior, or the prior needs to include rules for how to determine the match of a whole utterance on the basis of the match with each component. In a later experiment of Degen et al. (2020), typicality effects made compositional semantics not work, but the utterance space was small enough that each utterance could be treated individually.

In less toy domains, there is not a satisfactory answer: some situations can be handled by empirically measuring likelihoods in an exhaustive ways, but this holistic approach is not compatible with incremental RSA (Cohn-Gordon, Goodman, et al., 2018) or larger sets of utterances that require compositionality to be defined. Some work has used grounded neural language models for alternative generation and likelihood measures, which has scaling potential in domains that have the right grounded datasets (Cohn-Gordon, Goodman, et al., 2018; Monroe & Potts, 2015; White et al., 2020). In order to extend RSA towards more realistic and open-ended scenarios, an important question to grapple with is what form of meaning (even at a computational level) will appropriately support pragmatic reasoning.

4.2 RSA approaches to reduction

Scaling up RSA to handle reference games requires solving at least two problems. One is specifying a semantics system that can handle the abstract, metaphoric, and parts-based descriptions that are used – this could be seen as the problem of accounting for initial reference.

Secondly, one must explain how to go from one successful utterance to a different (often shorter) utterance. One RSA-style model that attempts to explain why multi-part descriptions are produced initially, but shorter descriptions are produced later is CHAI, a framework to bridge different levels of convention formation (R. D. Hawkins, Franke, et al., 2021).

CHAI incorporates parameterized hierarchical uncertainty over lexica that allow for variability in how words will be interpreted that can be integrated over. Listeners' lexica are not fully known which accounts for background differences that may be shared by a sub-population and differences from communication history that may be individual-specific. The results of inference update these priors, and propagate the new information to both the individual and group lexical representations. CHAI can qualitatively account for various convention-formation phenomena in toy systems (R. D. Hawkins, Franke, et al., 2021).

CHAI's explanation for reduction is consistent with the ambiguity arguments of Piantadosi et al. (2012): initially the context is not sufficiently constraining to resolve the ambiguity of a short utterance, so multiple descriptive pieces are needed to triangulate the meaning, but as conversational history accrues, the context is sufficient to disambiguate the shorter description.

In toy models of interlocuters playing a reference game with soft semantics, initial utterances use multiple properties to collectively increase the degree of certainty in the target. This successful reference then shapes the prior over word meanings, until the degree of certainty afforded by only one word is sufficient to pick out a referent.

The CHAI framework seems promising in that it qualitatively produces some of the patterns that have been most difficult to explain, namely why it would make sense to change descriptions when nothing in the context has changed. CHAI identifies that something in the context has changed – the conversation history has evolved and changed the word-meaning pairings in the speaker and listener lexica. This solves the issue of presenting a plausible and somewhat testable theory. It leaves open a lot of implementation questions about how to scale it up to be able to interact with (non-simulated, open vocabulary) data, and the model is likely to have a fair number of free parameters so it's unclear how to test it stringently.

4.3 Takeaways

RSA seems like the most relevant theoretical framework to a sequential (level 2 in the taxonomy in the beginning) model of iterated reference games. Two potential problems are scaling up to an open vocabulary and free parameters that will lead to high flexibility and a lack of risky prediction (Meehl, 1990). The open vocabulary problem is a hard one; rather than wait for a fully realized semantic system, I think one criteria for judging the adequacy of semantic systems is whether they can serve as a linking hypothesis to allow RSA models to predict the patterns of pragmatic language use that are observed experimentally. CHAI, while not a full explanation for the full patterns of real-world data, is a big step forward in at least providing a framework that actually explains why reduction would be optimal. It remains to be seen whether RSA-style models can predict the slope of reduction and the content of words that stay versus drop, but I believe the attempt would push forward our understanding of pragmatics and communication.

5 Psycholinguistic considerations

The utterances in reference games that optimization-oriented theories seek to explain are the product of lower-level, incremental processes. The algorithmic level of linguistic communication is constrained by its instantiation in the mind. To the extent we can infer these constraints from fine-grained behavior or transfer these constraints from other areas of psycho- and neuro-linguistics, these constraints may provide bounds for the computational models or provide testable predictions at these other levels of analysis.

5.1 Top-down or bottom-up

One large question in language processing and production broadly is what the relative balance of bottom-up and top-down influences are (Gwilliams et al., 2022; Horton & Gerrig, 2005; Horton & Keysar, 1996; Tanenhaus et al., 1995).

In the psycholinguistics of reference games, a hotly debated issue is whether the early moments of production and processing are "ego-centric" or can be influenced by non-linguistic information, such as the perspective of the interlocuter. On the production side, Horton and Keysar (1996) attempted to test this theory by comparing utterances produced with and without time pressure. They interpret the apparently ego-centric utterances in the speeded condition as evidence that initial utterance planning is ego-centric, but that monitoring and fixing of the utterance may take into account the listener's perspective, and may occur prior to utterance initiation.

On the comprehension side, Keysar et al. (2000) argued for an initially egocentric perspective on the basis that people often initially look at objects that are good matches to a description even if the object is not mutually visible to the speaker. This interpretation rests on a couple dubious linking hypotheses: that if people consider an interlocuters perspective, their prior should be that the interlocuter only refers to mutually known things; and separately, that looking at an object is a sign that it is considered a potential referent (eye-tracking data is widely interpreted this way, but there is evidence that this proxy is only approximate (Degen et al., 2021)).

The counterpoint to initial ego-centrism presented by Hanna et al. (2003) is a constraint-based theory where many factors can play into comprehension, including working memory limitations. Many factors may influence language production and comprehension, to varying extents, and on differing time courses. Determining the relative timings and weights is an important endeavor that will require careful experiments over a systematically varied swatch of experimental space.

As these experiments show, it's empirically difficult to do so when the measurements are far from the constructs and there are many nuisance variables to abstract over. An experiment compares two conditions with different objects visible to the speaker but not the listener (or vice versa), and can claim that for this set up and these objects, people look at the privileged object some amount. Is looking at the object predicted only by ego-centrism theory or can it also be predicted by some other pattern such as the constraint based approach? When a different experiment with different stimuli has a different result, is this because of nuisance differences in the stimuli? Or because of critical differences in how well each stimuli matched the verbal description? Even if we had well-characterized descriptive work about when this early-reaction occurs in terms of stimuli and conditions, there's still an interpretive question of what the early looks mean.

Some of the experiments around egocentrism try adding time pressure to get at earlier stages of production, but we don't actually know how time pressure interacts with the production system – does time pressure cause people to short-cut off the end, or to satisfice more at all the stages?

Perhaps more fully fleshed out versions of initial-ego-centrism theory and its alternatives and their requisite linking hypotheses could define a set of experiments where they each make clear and differing predictions. For now, with the existing theories and experiments, interpretations lie on a bunch of promises about assumed (untested) generality and linking assumptions.

5.2 Production constraints

In additional to possible information-integration timing limitations on optimality, the retrieval and generation of utterances may be another source of deviation from optimal models. Utterance planning is difficult, and production biases such as easy first, plan reuse, and reduce interference may produce deviations from information-theoretic predictions (MacDonald, 2013).

Another limitation is the search problem of production. RSA and other computational theories assume the existence of alternatives and then provide ways of choosing from among the options. Especially with low-codability targets, the initial mental generation of any potential referring expression may be a bottleneck. There are empirical challenges with determining what is difficult for speakers to produce, although analyses of disfluencies is one approach (S. O. Yoon & Brown-Schmidt, 2014). Getting traction on initial utterance planning could usefully inform the generation of alternatives in information-theoretic models, but this is likely very challenging.

5.3 The need for different kinds of data

From a production side, we might want to know what constraints are how influential on speakers utterances (for instance, wanting to know if they are initially egocentric). If all we have is the transcript of what is eventually produced, there's not much we can do. We'd have to make a lot of assumptions about how the time course of production maps to a final utterance. Maybe we can say something based on what types of phrases are produced at the start of an utterance versus the end. But this still assumes that the time scale of articulation reflects the timescale of planning. Which is often true, but assumes that speakers don't pause to fully plan their utterance before beginning to emit the utterance.

Timed data could address this by revealing where there are (filled or unfilled) pauses, and how long speakers take before initiating the utterance. This is still an indirect measure of what is happening in the mind, but it gives richer data for theories to fit to and is thus a stricter test.

We still can't know even from this data whether there's a very early initial stage that works somehow but is overcome before any utterance exits the mouth. Here, we would want other measures, such as possibly eye-tracking or neural data to look for evidence of what is being thought about prior to utterance initiation. As a caveat, interpreting these signals will still rest on a number of assumptions.

For the comprehension side, accuracy data provides the least constraint and insight into the process, reaction time provides a little more, and eye-tracking or other continuous measures may provide more, although they still require linking assumptions.

5.4 Takeaways

Eventually models and evidence at all levels need to converge for a satisfying network of theories. While many of the phenomena of interest are described at a fairly high level in terms of observed utterances, the mechanisms by which these utterances and interpretations are produced go through the language processing and production system.

In some cases, the questions of interest are about the fine-grained time course of reference games: how utterances are generated and interpreted and what information is integrated when. In other cases, the theories rely on implicit assumptions, such as ideas that the language system does something that can be approximated as generating and evaluating alternatives. These assumptions and questions are about the the general processes of language cognition which this instance of language use shares with other instances of language use.

Even if they are not the target, psycholinguistic considerations need to be part of the parsimony of theories. High-level theories write promissory notes that at least algorithmic approximations will be found for linguistic or memory processes – we should seek theories that make promises that are most compatible with psycholinguistic findings and theories.

6 Ways forward

A key part of formal theories and computational models is they need to make clear quantitative predictions that can account for the existing data regarding the phenomena of interest.

Before we can do that, we need is a better definition of what is to be predicted. I believe the features of central interest here are a) how people describe images/objects/things where there is not a canonical or conventional name and b) the dynamics of how this process, over repetition, results in nicknames.

The gap between data and theory can be narrowed from both the data and theory sides, so I lay out a few potential avenues for progress.

6.1 Data side

One necessity is a better empirical and descriptive understanding of the phenomena of interest. How do circumstantial knobs (group size, communication modality, stimuli, etc.) push around the extend to which we observe reduction, convention formation, and partner specificity?

We don't have to finely map the entire landscape, but some sense of the geography of experimental space is necessary (Almaatouq et al., 2022). Secondly, with these related phenomena, there's a question of how closely they are tied. Does reduction only occur via convention formation? A greater range of experimental situations could tell how well these separate versus correlate in different situations, which will in turn clarify what theory should explain.

Another empirical avenue is to fill in descriptive details on the process of reduction. This will require finding useful joints to carve the rich language data into. Quantitatively, what types of utterances tend to occur after what other utterances? One could imagine trying to map the probabilistic FSA and transition probabilities that descriptively characterize language of iterated reference games.

Another third approach is to push down a level of analysis and do empirical work looking at how utterances are produced rather than just recording what descriptions are produced. For initial descriptions, speakers have a hard task of generating a description. We can think of this as having a relatively flat distribution over options, where no words have a particularly strong salience (no previous rounds to go on) and nothing has an especially high fit because of the low codeability. The combined flatness means that generating an entire description is particularly difficult to predict, but understanding the incremental process may be more tractable.

6.2 Theory side

In the introduction, I listed three scales on which one might try to make theories: the scale of an entire conversation, the scale of one instance of referring, and the scale of incremental production and comprehension. I think the most viable of these is the second, where we try to get predictive or causal traction of the incremental steps in the conversation.

This scale is the closest to having formal models available, in the form of RSA-style models and CHAI. The conversation history up to the point of the utterance can be used to shape the prior, by updating a baseline prior to make previous utterances more available and increasing the weight on meaning-utterance pairings that were evidenced in the prior rounds. Then the goal of the model would be either (speaker-side) predicting the next utterance or (listener-side) predicting how the prior utterance will be interpreted, which could be put thought a linking function to result predicting either a guess or a request for more information. This parallels the RSA task of predicting an utterance option or predicting an utterance interpretation depending on whether it is listener or speaker side.

Two big questions here are how to handle the open-ended semantics and how to construct the alternative set, as discussed above. In addition to trying to scale to match the open-ended semantics and alternatives, one could instead switch the prediction problem to a different scale of granularity. One possibility is to predict not the lexical items themselves, but some of the properties, such as the number of clauses, or the semantic attributes. Thus, rather than needing to predict whether a speaker says "ice skater" or "ice skater with their arms up" or "ice skater with a leg out to the left" or "t-rex" or "figure with one arm up and one leg out", one could try to code this in terms of "holistic description – new" versus "holistic description used in first round" versus "body part" and then for each target and description-type pair specify an initial semantics, combination rules, and update rules. This sort of coarser approach may be more tractable, but also has lower predictive value and bakes in more assumptions. A question here is what would constitute a viable minimal non-toy model of the reduction and partner-specificity phenomena.

Another direction where RSA models don't have the capacity to match the type of data from reference games is for non-dyadic reference games. Group communication, either in the form of one speaker - multiple listeners (S. O. Yoon & Brown-Schmidt, 2014, 2018; S. O. Yoon & Brown-Schmidt, 2019a) or network structures (R. D. Hawkins, Liu, et al., 2021), poses a challenge as there is not yet a non-dyadic extension of RSA that can do partial pooling across listeners and jointly optimize for the responses of multiple interlocuters.

The other main theoretical approach is the efficiency angle. Unfortunately, efficiency makes vague predictions and may be hard to crystallize out due to the flexibility of many possible linking assumptions. One approach would be to treat the transcripts as some sort of artifact and try to determine potential counterfactuals, to see where the attested patterns fall on some larger space. Defining the space of possibilities could be difficult, and determining the information value of counterfactuals may require empirical testing. But one could address the open question of whether on a comprehension side, the observed patterns are efficient. This leaves out the potential that production side is the bottleneck on efficiency. Testing efficiency in this way would not definitively answer "are people efficient?", but could narrow in on what assumptions are required to claim that people are efficient.

Current statistical models used to assess reduction (or disfluency rate, etc) have the issue that they are fit to a specific dataset and can't be compared across datasets and situations, so it's difficult to assess their quantitative generality (Yarkoni & Westfall, 2017).

7 Outline of the dissertation

In this dissertation, I focus on starting to fill in some of the data gaps discussed above. I hope that this work better characterizing the phenomena of interest across contexts and levels of analysis provides some ground-work for future theory development.

1) In the first substantive chapter, I take a broader look at when reduction and semantic convergence patterns occur by looking at larger groups and different modalities.

2) Using the descriptions from 1), I test some of the partner-specificity and efficiency-claims in a listener-only paradigm to see how naive listeners understand descriptions from various timepoints in games. Additionally, I use incremental methods to understand some of the time course of comprehension of these descriptions.

3) Finally, I examine some of the developmental origins of the skills behind ad-hoc reference with experimental results from 4-5 yo children playing a scaled down version of iterated reference games.

References

Abbot-Smith, K., Nurmsoo, E., Croll, R., Ferguson, H., & Forrester, M. (2016). How children aged 2;6 tailor verbal expressions to interlocutor informational needs. *Journal of Child Language*, *43*(6), 1277–1291. https://doi.org/10.1017/S0305000915000616

Almaatouq, A., Griffiths, T. L., Suchow, J. W., Whiting, M. E., Evans, J., & Watts, D. J. (2022). Beyond Playing 20 Questions with Nature: Integrative Experiment Design in the Social and Behavioral Sciences. *Behavioral and Brain Sciences*, 1–55. https://doi.org/10.1017/S0140525X22002874

Bergen, L., Levy, R., & Goodman, N. (2016). Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, 84.

Branigan, H. P., Bell, J., & McLean, J. F. (2016). Do You Know What I Know? The Impact of Participant Role in Children's Referential Communication. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00213

Brennan, S. E., & Clark, H. H. (1996). Conceptual Pacts and Lexical Choice in Conversation, 12.

Brown-Schmidt, S. (2012). Beyond common and privileged: Gradient representations of common ground in real-time language use. *Language and Cognitive Processes*, *27*(1), 62–89. https://doi.org/10.1080/01690965.2010.543363

Clark, H. H. (1996). *Using Language*. Retrieved August 29, 2023, from https://www.cambridge.org/core/books/using-language/4E7EBC4EC742C26436F6CF187C43F239

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*. Retrieved October 6, 2020, from http://www.speech.kth.se/~edlund/bielefeld/references/clark-and-wilkes-gibbs-1986.pdf

Cohn-Gordon, R., Goodman, N., & Potts, C. (2018, May 10). *Pragmatically Informative Image Captioning with Character-Level Inference*. Retrieved November 6, 2021, from http://arxiv.org/abs/1804.05417

Cohn-Gordon, R., Goodman, N. D., & Potts, C. (2018, October 19). *An Incremental Iterated Response Model of Pragmatics*. Retrieved May 4, 2020, from http://arxiv.org/abs/1810.00367

Degen, J., Hawkins, R. D., Graf, C., Kreiss, E., & Goodman, N. D. (2020). When redundancy is useful: A Bayesian approach to "overinformative" referring expressions. *Psychological Review*, *127*(4), 591. https://doi.org/10.1037/rev0000186

Degen, J., Kursat, L., & Leigh, D. (2021). Seeing is believing: Testing an explicit linking assumption for visual world eye-tracking in psycholinguistics. *CogSci*.

Fay, N., Garrod, S., Roberts, L., & Swoboda, N. (2010). The Interactive Evolution of Human Communication Systems. *Cognitive Science*, *34*(3), 351–386. https://doi.org/10.1111/j.1551-6709.2009.01090.x

Ferreira, F. (2004). Production-comprehension asymmetries. *Behavioral and Brain Sciences*, *27*(2), 196–196. https://doi.org/10.1017/S0140525X04280050

Fox Tree, J. E., & Clark, N. B. (2013). Communicative Effectiveness of Written Versus Spoken Feedback. *Discourse Processes*, *50*(5), 339–359. https://doi.org/10.1080/0163853X.2013.797241

Frank, M. C., & Goodman, N. D. (2012). Predicting Pragmatic Reasoning in Language Games. *Science*, *336*(6084), 998–998. https://doi.org/10.1126/science.1218633

Franke, M., & Degen, J. (2016). Reasoning in Reference Games: Individual- vs. Population-Level Probabilistic Modeling (P. Allen, Ed.). *PLOS ONE*, *11*(5), e0154854. https://doi.org/10.1371/journal.pone.0154854

Gandolfi, G., Pickering, M. J., & Garrod, S. (2022). Mechanisms of alignment: Shared control, social cognition and metacognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *378*(1870), 20210362. https://doi.org/10.1098/rstb.2021.0362

Garrison, A. C. S., Yoon, S. O., Brown-Schmidt, S., Ariss, T., & Fairbairn, C. (2022, October 28). *Alcohol and Common Ground: The Effects of Intoxication on Linguistic Markers of Shared Understanding during Social Exchange*. https://doi.org/10.31219/osf.io/xrw6z

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How Efficiency Shapes Human Language. *Trends in Cognitive Sciences*, *23*(5), 389–407. https://doi.org/10.1016/j.tics.2019.02.003

Glucksberg, S., Krauss, R., & Weisburg, R. (1966). *Referential Communication in Nursery School Children: Method and Some Preliminary Findings.*

Glucksberg, S., & Krauss, R. M. (1967). WHAT DO PEOPLE SAY AFTER THEY HAVE LEARNED HOW TO TALK? STUDIES OF THE DEVELOPMENT OF REFERENTIAL COMMUNICATION. *Merrill-Palmer Quarterly of Behavior and Development*, *13*(4), 309–316. Retrieved June 6, 2022, from https://www.jstor.org/stable/23082551

Goodman, N. D., & Frank, M. C. (2016). Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences*, *20*(11), 818–829. https://doi.org/10.1016/j.tics.2016.08.005

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and Implicature: Modeling Language Understanding as Social Cognition. *Topics in Cognitive Science*, *5*(1), 173–184. https://doi.org/10.1111/tops.12007

Guilbeault, D., Baronchelli, A., & Centola, D. (2021). Experimental evidence for scale-induced category convergence across populations. *Nature Communications*, *12*(1), 327. https://doi.org/10.1038/s41467-020-20037-y

Gwilliams, L., Marantz, A., Poeppel, D., & King, J.-R. (2022). Top-down information flow drives lexical access when listening to continuous speech. *bioRxiv.*

Hanna, J. E., Tanenhaus, M. K., & Trueswell, J. C. (2003). The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language*, *49*(1), 43–61. https://doi.org/10.1016/S0749-596X(03)00022-6

Hawkins, J. (1995). A Performance Theory of Order and Constituency. *Cambridge University Press.*

Hawkins, R. D., Frank, M. C., & Goodman, N. D. (2020, April 13). *Characterizing the dynamics of learning in repeated reference games*. Retrieved July 15, 2020, from http://arxiv.org/abs/1912.07199

Hawkins, R. D., Franke, M., Frank, M. C., Goldberg, A. E., Smith, K., Griffiths, T. L., & Goodman, N. D. (2021, December 2). *From partners to populations: A hierarchical Bayesian account* *of coordination and convention.* arXiv: 2104.05857 [cs]. https://doi.org/10.48550/arXiv.2104.05857

Hawkins, R. D., Liu, I., Goldberg, A. E., & Griffiths, T. G. (2021). Respect the code: Speakers expect novel conventions to generalize within but not across social group boundaries. *CogSci.*

Horton, W. S., & Gerrig, R. J. (2002). Speakers' experiences and audience design: Knowing when and knowing how to adjust utterances to addressees. *Journal of Memory and Language*, *47*(4), 589–606. https://doi.org/10.1016/S0749-596X(02)00019-0

Horton, W. S., & Gerrig, R. J. (2005). The impact of memory demands on audience design during language production. *Cognition*, *96*(2), 127–142. https://doi.org/10.1016/j.cognition.2004.07.001

Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition*, *59*(1), 91–117. https://doi.org/10.1016/0010-0277(96)81418-1

Ibarra, A., & Tanenhaus, M. K. (2016). The Flexibility of Conceptual Pacts: Referring Expressions Dynamically Shift to Accommodate New Conceptualizations. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00561

Kao, J. T. (2014). Formalizing the Pragmatics of Metaphor Understanding. *CogSci.*

Kemp, C., Xu, Y., & Regier, T. (2018). Semantic Typology and Efficient Communication. *Annual Review of Linguistics*, *4*(1), 109–128. https://doi.org/10.1146/annurev-linguistics-011817-045406

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking Perspective in Conversation: The Role of Mutual Knowledge in Comprehension. *Psychological Science*, *11*(1), 32–38. https://doi.org/10.1111/1467-9280.00211

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102. https://doi.org/10.1016/j.cognition.2015.03.016

Köymen, B., Schmerse, D., Lieven, E., & Tomasello, M. (2014). Young children create partner-specific referential pacts with peers. *Developmental Psychology*, *50*(10), 2334–2342. https://doi.org/10.1037/a0037837

Krauss, R. M., & Weinheimer, S. (1966). Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology*, *4*(3), 343–346. https://doi.org/10.1037/h0023705

Leung, A., Yurovsky, D., & Hawkins, R. (2023, April 1). *Parents scaffold the formation of conversational pacts with their children.* https://doi.org/10.31234/osf.io/8u4qa

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00226

Matthews, D., Lieven, E., & Tomasello, M. (2010). What's in a manner of speaking? Children's sensitivity to partner-specific referential precedents. *Developmental Psychology*, *46*(4), 749–760. https://doi.org/10.1037/a0019657

Meehl, P. E. (1990). Appraising and Amending Theories: The Strategy of Lakatosian Defense and Two Principles that Warrant It. *Psychological Inquiry*, *1*(2), 108–141. https://doi.org/10.1207/s15327965pli0102_1

Metzing, C., & Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, *49*(2), 201–213. https://doi.org/10.1016/S0749-596X(03)00028-7

Monroe, W., & Potts, C. (2015, October 22). *Learning in the Rational Speech Acts Model.* Retrieved September 24, 2020, from http://arxiv.org/abs/1510.06807

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, *21*(5), 1112–1130. https://doi.org/10.3758/s13423-014-0585-6

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280–291. https://doi.org/10.1016/j.cognition.2011.10.004

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(02). https://doi.org/10.1017/S0140525X04000056

Qing, C., Goodman, N. D., & Lassiter, D. (2016). A rational speech-act model of projective content. *CogSci*.

Rakoczy, H. (2022). Foundations of theory of mind and its development in early childhood. *Nature Reviews Psychology*, *1*(4), 223–235. https://doi.org/10.1038/s44159-022-00037-z

Rogers, S. L., Fay, N., & Maybery, M. (2013). Audience Design through Social Interaction during Group Discussion. *PLOS ONE*, *8*(2), e57211. https://doi.org/10.1371/journal.pone.0057211

Rubio-Fernandez, P., Mollica, F., & Jara-Ettinger, J. (2021). Speakers and listeners exploit word order for communicative efficiency: A cross-linguistic investigation. *Journal of Experimental Psychology: General*, *150*, 583–594. https://doi.org/10.1037/xge0000963

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634. https://doi.org/10.1126/science.7777863

White, J., Mu, J., & Goodman, N. D. (2020, May 30). *Learning to refer informatively by amortizing pragmatic reasoning*. Retrieved November 6, 2021, from http://arxiv.org/abs/2006.00418

Wilkes-Gibbs, D., & Clark, H. (1992). Coordinating beliefs in conversation. *Journal of Memory and Language*, 183–194.

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Yoon, E. J., Frank, M. C., Tessler, M. H., & Goodman, N. D. (2018). Polite speech emerges from competing social goals. https://doi.org/10.31234/osf.io/67ne8

Yoon, S. O., & Brown-Schmidt, S. (2014). Adjusting conceptual pacts in three-party conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(4), 919–937. https://doi.org/10.1037/a0036161

Yoon, S. O., & Brown-Schmidt, S. (2018). Aim Low: Mechanisms of Audience Design in Multiparty Conversation. *Discourse Processes*, *55*(7), 566–592. https://doi.org/10.1080/0163853X.2017.1286225

Yoon, S. O., & Brown-Schmidt, S. (2019a). Audience Design in Multiparty Conversation. *Cognitive Science*, *43*(8), e12774. https://doi.org/10.1111/cogs.12774

Yoon, S. O., & Brown-Schmidt, S. (2019b). Contextual Integration in Multiparty Audience Design. *Cognitive Science*, *43*(12), e12807. https://doi.org/10.1111/cogs.12807

Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, *115*(31), 7937–7942. https://doi.org/10.1073/pnas.1800521115

Zipf, G. (1949). *Human behavior and the principle of least effort: An introduction to human ecology.*