THE FUTURE OF SCHOLARSHIP

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF PSYCHOLOGY
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Mary J. Stanford

February 2025

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Advisor T. Greatest) Principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(L. O. Sunshine)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Carl Friedrich Gauss)

Approved for the Stanford University Committee on Graduate Studies

Abstract

I'm an abstract right here, look at me!

Dedication

Acknowledgments

Contents

List of Tables

List of Figures

Chapter 1

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

%TODO FOR LATER an example can strengthen this kind of abstract info and make it m

Language is an amazing human technology that supports much of human culture by bei

Much linguistic creativity is ephemeral, occurring at at a single time within a si

Whether descriptions are fleeting or lasting, the core observation is the same: hu

How can we explain and predict these phenomena? Following \citet{meehl1990], we wi

%TODO FOR LATER can you say a bit about why we'd want to do this? other than "beca

%TODO FOR LATER be more explicit about the Philosophy of science backing here

1.1 Levels of explanation

One issue with developing formal models of human linguistic behavior is that the data ar

To ground this, the common experimental paradigm for exploring communication about non-c

One can define the question of interest, and thus the scope of the modeling and predicti

1. A clear trend observed in the iterated reference games is that utterance length reduc

2. Instead of predicting the entire curve of a conversation, we could instead ask: given

3. Finally, one may wish to model the actual production and comprehension processes. We

%TODO FOR LATER Mike would like a diagram or table or **something**

Currently, we have formal models at none of these levels. Personally, I find the 2nd lev

At whatever level, we would like to eventually have theories that are precise in wha

I will cover approaches from four perspectives. First, the communication and reference g

1.2 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.

%The communication and reference game literature provides descriptive theories that identify some phenomena of interest and raise questions around whether these phenomena occur intentionally or emerge as a by-product of other processes. These serve as a list of results that models could hope to explain.

%However, our 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.

1.2.1 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.

%TODO FOR LATER running example should reoccur here!

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 & Wilkes-Gibbs, 1986). 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 & Wilkes-Gibbs, 1986; 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.

%TODO FOR LATER Gabe Doyle stuff

%TODO FOR LATER connect to the lexical side 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]. %TODO FOR 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] TODO FOR LATER 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 \citet{krauss1966]) 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.

1.2.2 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; 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 (**hawkins2020b?**). 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 (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, 2002b, 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 & Wilkes-Gibbs, 1986; 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, 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.

1.2.3 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. \citet{horton2002a] 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 (**hawkins2020b?**). Thus “reduction” or the shortening of utterances is sometimes used as a shorthand and measurement proxy for the semantic changes (Clark & Wilkes-Gibbs, 1986; 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.

1.2.4 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 & 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 H. P. 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 & 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.

1.2.5 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.

%One sign here may be that it would be difficult to meta-analyze the results to look for moderators like group size; for instance, for reduction phenomena, should it be average decrease in words per repetition? percent decrease in words from first to second repetition? There's hints that we think there's some functional form where the rate of decrease should be higher the longer the utterance is and there's an expected plateau at some floor

1.3 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.

1.3.1 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.

1.3.2 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; **hawkins2020b?**). "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.

1.4 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.

1.4.1 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 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 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.

%TODO FOR LATER re-read the neural RSA stuff

1.4.2 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 (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 (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.

1.4.3 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.

1.5 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.

%TODO FOR LATER mike is dubious about this for so-called “computational level” models or something

%TODO FOR LATER maybe cite the new shain paper as an example of how different levels influence each other

1.5.1 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 & 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.

1.5.2 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.

1.5.3 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.

1.5.4 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.

1.6 Ways forward

%What would we want from formal theories and computational models?

%Any theory also needs to be compatible with what we know at different levels of analysis and related phenomena. A theory of linguistic iterated reference to novel objects need not cover how named objects are referred to, or how conventions form in drawing, or how language evolves, or how language is processed in the brain. But it is important the theory be compatible with all of these; it should not make untrue assumptions about these other areas. Ideally, the boundaries of a theory would be continuous with adjacent results and theories.

%Related, there are questions of how to scope a theory of iterated reference games; for instance, it may make sense to have a model that only covers dyadic interactions or a certain modality of interaction or a certain type of stimuli. This limited scope is fine if it is clear and the assumptions of the models don't lead to wild (false) predictions.

A key part of formal theories and computational models is they need to make clear

Before we can do that, we need is a better definition of what is to be predicted.

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.

1.6.1 Data side

One necessity is a better empirical and descriptive understanding of the phenomena of in

We don't have to finely map the entire landscape, but some sense of the geography of exp

Another empirical avenue is to fill in descriptive details on the process of reduction.

Another third approach is to push down a level of analysis and do empirical work looking

1.6.2 Theory side

In the introduction, I listed three scales on which one might try to make theories%TODO
: the scale of an entire conversation, the scale of one instance of referring, and the s
I think the most viable of these is the second, where we try to get predictive or causal

This scale is the closest to having formal models available, in the form of RSA-style mo

%Basically, given a history to a certain point, how to predict what interpretations or u
Two big questions here are how to handle the open-ended semantics and how to construct t

%TODO FOR LATER ugh this is getting away from RSA though
%The big questions here are how to handle open-ended semantics and how to construct the

Another direction where RSA models don't have the capacity to match the type of data fro

The other main theoretical approach is the efficiency angle. Unfortunately, efficiency m

Current statistical models used to assess reduction (or disfluency rate, etc) have t

1.7 Outline of the dissertation

%TODO FOR LATER FOR LATER WHAT'S THE CAUSAL MODEL SO I CAN HAVE A CUTE OVERVIEW DI

In this dissertation, I focus on starting to fill in some of the data gaps discuss

1) In the first substantive chapter, I take a broader look at when reduction and s

% 2) Using the transcripts from 1), I then dig into the verbal data at a finer-grained level to get a grasp on at the more utterance-to-utterance evolution of descriptions.

2) Using the descriptions from 1), I test some of the partner-specificity and effi

3) Finally, I examine some of the developmental origins of the skills behind ad-ho

%

% Hawkins et al. (2020) How do you break symmetries in initial descriptions: there's prior variability across speaker preferences (they may each have a preferred label, but be unsure if others will accept it) and/or speakers may also not have labels and need to do some sampling. This doesn't account for production time course factors. % % Hawkins, Franke, et al. (2021) says there are 3 core cognitive abilities: the ability to represent that there is variability in other's lexicons; to coordinate via online learning; and to generalize across interactions ("partial pooling" model where updates both to partner and population) % %Efficiency also predicts that a changing conversational history will change the context and thus different descriptions may be efficient. This could operate both by increasing beliefs that a certain utterance will be understood (and this is contextually low ambiguity) or more generally by shaping the syntactic expectations, perhaps making it easier to produce and comprehend odder descriptions. % %Hawkins, Franke, et al. (2021) separating the inference problem about what the other person's lexicon is (which is how they will interpret things in the moment, b/c you may have temporarily changed their lexicon) with decisions about what to say given that % %Hawkins, Franke, et al. (2021) points out that our models for communication and modeling the world and others need to be able to account for different people having different knowledge (including some tied to community membership or social role) and that vocabularies need to accomodate change over time and new things to refer to as the world changes.

%TODO FOR LATER FIX REFERENCES!!!

Chapter 2

Interaction structure

2.1 Introduction

Much of human social life revolves around communication in groups. At school, teachers address large classrooms of children (Cazden, 1988); at home, we chat with groups of friends and family members over dinner (Tannen, 2005); and at work, we attend meetings with colleagues and managers (Caplow, 1957; Zack, 1993). Such settings present considerable challenges that do not arise in the purely two-party (dyadic) settings typically studied in psychology (H. Branigan, 2006; Ginzburg & Fernandez, 2005; Traum, 2004). For example, producers need to account for the fact that different comprehenders in the group may have different mental states or levels of background understanding (Fox Tree & Clark, 2013; Horton & Gerrig, 2002a; Horton & Gerrig, 2005; Weber & Camerer, 2003; S. O. Yoon & Brown-Schmidt, 2014, 2018), while comprehenders must account for the fact that utterances are not necessarily tailored to them (Carletta et al., 1998; Cohn-Gordon et al., 2019; Fay et al., 2000; Metzing & Brennan, 2003; Rogers et al., 2013; Tolins & Fox Tree, 2016; S. O. Yoon & Brown-Schmidt, 2019a). What enables producers and comprehenders to nevertheless overcome these challenges and navigate multi-party settings with relative ease?

One promising set of hypotheses centers on the group's *interaction structure*, the set of constraints placed on the group's shared communication channel. Many different aspects of interaction structure have been implicated in the effectiveness of dyadic communication, including the availability and quality of concurrent feedback (Krauss

& Bricker, 1967; Krauss & Weinheimer, 1966; Kraut et al., 1982), the bandwidth of the communication modality (Dewhirst, 1971; Krauss et al., 1977), and the group's access to a shared workspace (Clark & Krych, 2004; Garrod et al., 2007). Yet larger groups introduce qualitatively different dimensions of interaction structure, leading to a large but often inconsistent body of findings even for these well-understood factors (Hiltz et al., 1986; Swaab et al., 2012). While communication is generally expected to deteriorate as groups get larger (MacMillan et al., 2004; Seaman & Basili, 1997), the structural "thickness" of the feedback channel may slow such deterioration (Ahern, 1994; Parisi & Brungart, 2005).

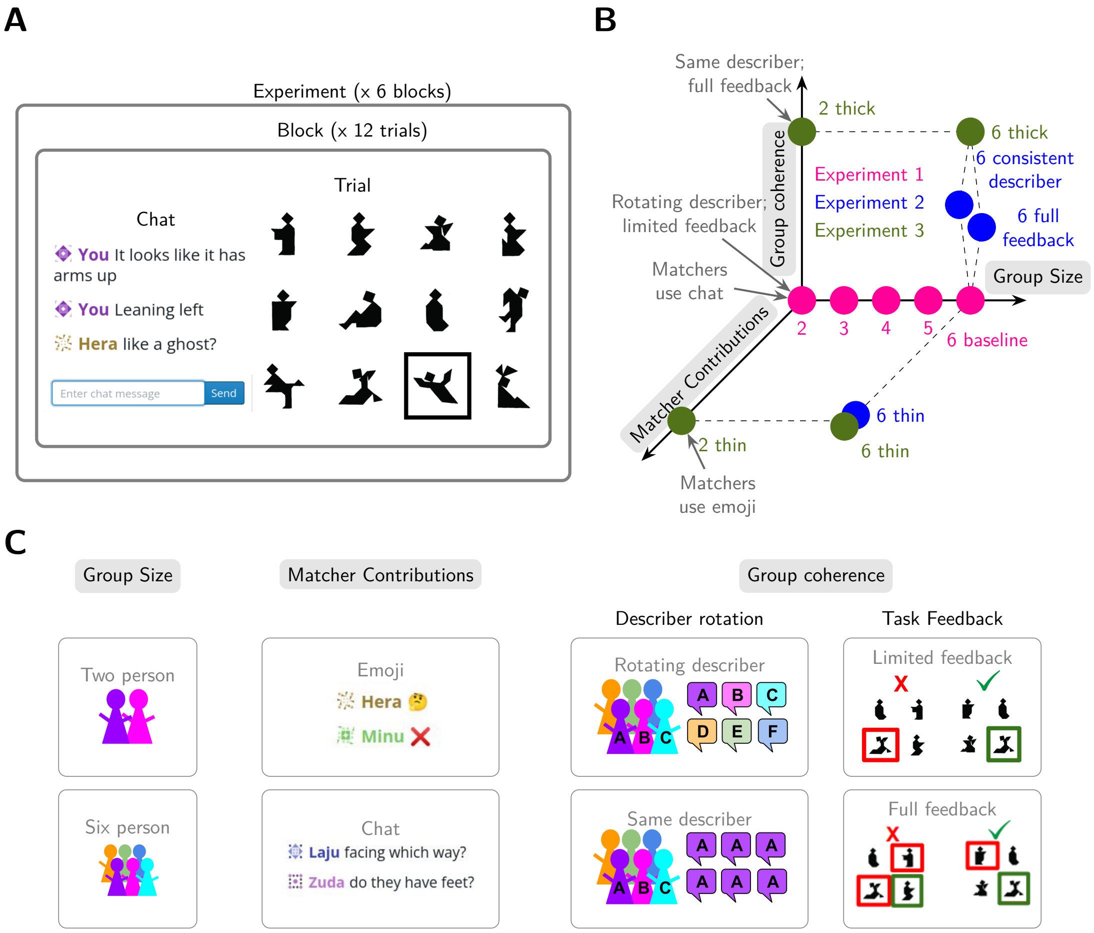
In this paper, we develop an experimental paradigm for evaluating the relative contribution of these factors: a *multi-party repeated reference game.* The ability to distinguish one particular entity from other possible entities, known as *reference*, is one of the most primitive and ubiquitous functions of communication. Reference games (Lewis, 1969; Wittgenstein, 1953) have been widely used to study dyadic communication under controlled conditions in the lab. They provide a clear metric of communicative effectiveness: how many words are required before a matcher successfully chooses a target image from a context of distractors? *Repeated* reference games, where the same target images appear multiple times in succession, were introduced to examine how interlocutors establish shared reference in the absence of conventional labels (Clark & Wilkes-Gibbs, 1986; Krauss & Weinheimer, 1964). At the beginning of the game, long and costly descriptions are typically required to succeed. A key finding, however, is that dyads become increasingly efficient over the course of interaction. Fewer words are required to achieve the same accuracy, but referring expressions also become more impenetrable to outsiders (Schober & Clark, 1989; **wilkes1992?**). The evolution of referring expressions over repetitions shows the characteristic dynamics of conventions: *stability*, or convergence on labels within a group, and *arbitrariness*, or divergence to different across groups, suggesting that dyads leverage their shared communication history to coordinate on expectations about how to label the target images (Hawkins et al., 2023).

In principle, repeated reference games provide a strong operationalization of communicative effectiveness for the problem of multi-party communication: describers must simultaneously achieve shared reference with multiple matchers. However, empirically studying multi-party communication raises a number of difficulties in practice. A much larger pool of participants must be recruited to achieve sufficient power at the relevant unit of analysis – the group – spanning a very high-dimensional space of possible parameter settings (Almaatouq et al., 2024). We address this problem by drawing on recent technical advances that have made it newly possible to achieve such samples using interactive web-based platforms (Almaatouq et al., 2020; Haber et al., 2019; Hawkins et al., 2023). Repeated reference games in web-based platforms have previously replicated earlier results from face-to-face studies, although people produce fewer words in text modalities than oral modalities (Hawkins et al., 2020). The text-based chat modalities arguably more closely resemble the interfaces used by modern teams who increasingly communicate through group text threads or popular platforms like Slack or Discord.

We leverage our platform to explore effects of group size and interaction channel thickness in a series of three experiments. While we find that small groups reliably converge on group-specific "shorthand" regardless of the interaction structure, larger groups require thicker channels – richer conversational feedback among members – to achieve the same degree of coherence. Thus, increasing group size alone does not impede communication; rather, larger groups may require stronger social and linguistic cues to establish common ground among all members. More broadly, our work suggests that studying communication in larger groups is necessary to unveil critical aspects of interaction structure that have not been evident in typical dyadic settings.

2.2 Results

We recruited 1319 participants through Prolific, an online crowd-sourcing platform. Participants were organized into 313 groups of size two to six for a communication game (Figure 2.2A). On each trial, everyone in the group was shown a gallery of 12 tangram images (Clark & Wilkes-Gibbs, 1986; Hawkins et al., 2020; **ji2022abstract?**). One player was designated the *describer* and the others were designated the *matchers*. The describer was asked to use a chat box interface to describe a privately indicated



You It looks like it has arms up

You Leaning left

Hera like a ghost?

Hera 🤔

Minu ❌

Laju facing which way?

Zuda do they have feet?

Figure 2.1: (A) Participants played a repeated reference game in groups of size 2 to 6. On each trial, a describer described the target image to the group of matchers. Each image appeared once per block for six blocks. (B) Experiments varied along 3 dimensions: Group size, group coherence, and matcher contributions. (C) Experiment 1 (pink) varied group size from 2 to 6 players while holding group coherence and matcher contributions constant. Experiment 2 (blue) held group size constant at 6 and manipulated the other dimensions. Experiment 3 (green) tested 4 corners of the space, crossing group size (2 vs. 6 players) with the thickness of interaction structure (high vs. low coherence and matcher contributions).

target image. After all matchers guessed which of the 12 images was the target, they received task feedback and proceeded to the next trial. The game consisted of 72 trials structured into 6 repetition blocks, where each image appeared as the target exactly once per block.

We manipulated the interaction structure of this game across 11 distinct conditions in 3 distinct pre-registered experiments (Figure 2.2B). We systematically sampled points along four dimensions parameterizing different aspects of the interaction space. We manipulated *group size* (ranging from two to six), *role stability* (whether or not participants took turns in the describer role), richness of *task feedback* (whether or not matchers were able to see each other's responses), and richness of the *matcher contributions* (whether matchers were able to freely respond through a chatbox or could only use emojis; Figure 2.2C). Other factors, such as the set of stimuli and background knowledge about one's partners, were held constant across games.

2.2.1 Overview of experiments

Experiment 1 began by investigating how performance scaled with group size. Based on prior qualitative work, we predicted that larger groups face a more challenging coordination problem. We continuously varied the number of players from 2 to 6 while keeping other factors constant. For these conditions, the describer role rotated after each block, so that all players had at least one turn as describer. Matchers had access to an unrestricted chat box, but only received binary task feedback about whether their individual selection was correct without revealing others' selections or the intended target.

Experiment 2 focused on the most challenging 6-player groups and explored the role of interaction structure. Each condition in Experiment 2 varied one aspect of the experiment relative to the Experiment 1 6-player baseline. We tried two variants that we expected to increase group coherence and improve performance, and a third variant we expected to interfere with the ability to establish mutual understanding and thus impede performance. In the first variant, we maintained the same describer throughout rather than a rotating describer, such that the same individual has the opportunity to aggregate feedback across trials and track which matchers are struggling with which targets. In the second variant, we gave the group of matchers full feedback about what every other member of the group had selected, and we showed the intended target. In the third variant, we changed how matchers could make contributions to the group. In contrast to prior experiments, where matchers could contribute freely to the chat; here, we limited matchers to sending four discrete emojis (green check, thinking face, red x, and laughing-crying face) that could convey simple valence and level of comprehension, but not any referential content.

Experiment 3 crossed the extremes of group size from experiment 1 (2 vs. 6 people) with the extremes of group interactions from Experiment 2 (*thick* vs. *thin* interaction structure). In the *thick* condition, we maintained a consistent describer, gave all matchers full task feedback, and allowed them to freely use a chat box. In the *thin* condition, we forced the describer to rotate on each block, restricted feedback to their own binary correctness, and restricted matcher contributions to the four emojis. Note that the 2-player thick game most closely resembles the design of classic repeated reference games (Clark & Wilkes-Gibbs, 1986).

2.2.2 Smaller and higher-coherence groups are more accurate

Our first set of hypotheses focused on group performance: how accurately and efficiently groups were able to perform the referential task. We characterize group performance along two complementary metrics: (1) matcher accuracy and (2) describer efficiency. Matcher accuracy is given by the percent of matchers on each trial who successfully selected the target referent. Describer efficiency is given by the number of words produced by the describer to achieve that degree of matcher accuracy in the group. The degree to which describers are able to communicate more efficiently without negatively impacting matcher accuracy is indicative of convergence on a more effective shared communication protocol within the group.

We begin by examining matcher accuracy, the extent to which the intended target was reliably transmitted to all matchers. We constructed a series of 5 logistic mixed-effects regression models predicting accuracy as a function of condition and repetition block (separate models were run for experiment 1, each condition in experiment 2, and experiment 3). For this and other effects, there was substantial variation at the tangram and game levels, with some tangrams being markedly easier than others and some groups performing differently than others. This wide variation made it difficult to precisely estimate population-level main effects, leading to wide credible intervals.

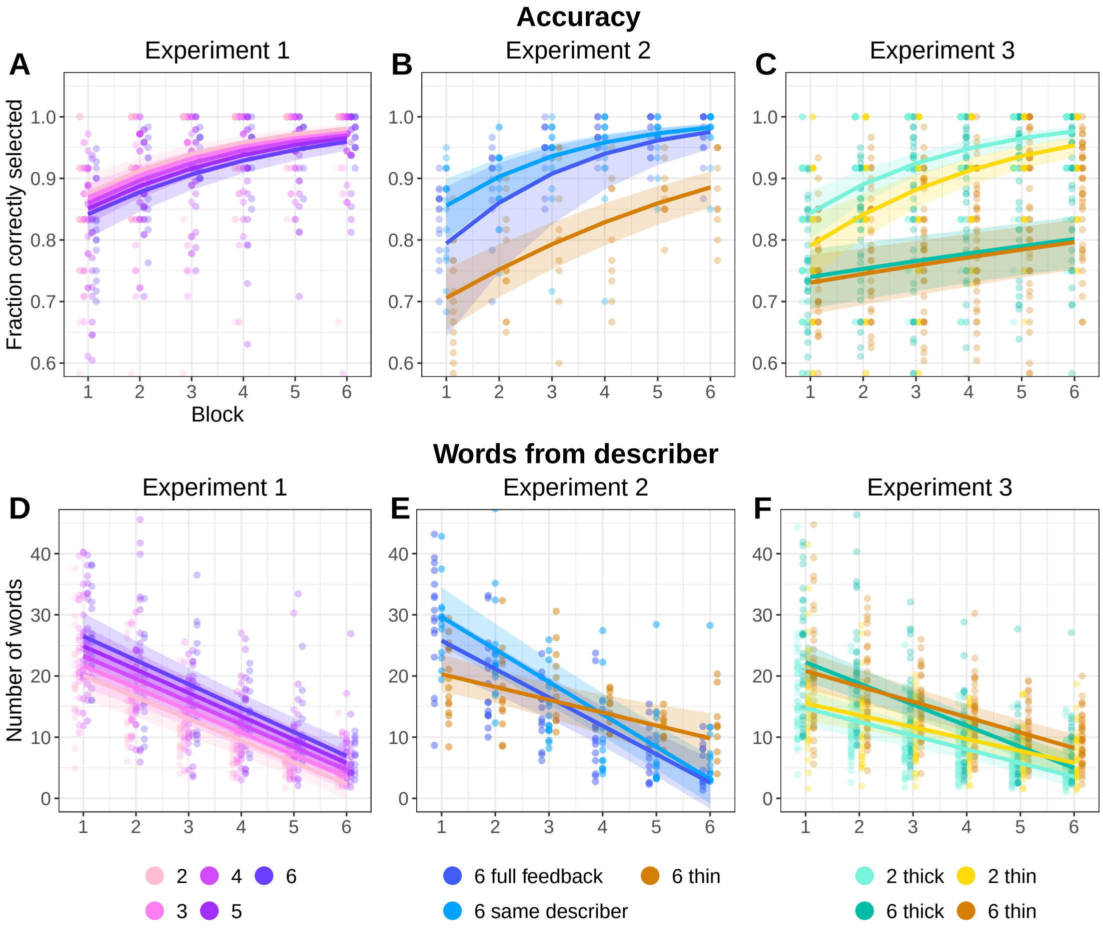


Figure 2.2: Behavioral results across all three experiments. (A-C). Matcher accuracy at selecting the target image. (D-F). Number of words produced by the describer each trial. For all, small dots are per game, per block means, and smooth lines are predictions from model fixed effects with 95% credible intervals. Y-axes are truncated, and a few outliers points are not visible.

See SI Figure 11 for a visualization of the relative magnitudes of population effects and game and tangram level variations.

Across all conditions, we observed strong positive effects of repetition block, indicating improved performance over time (Figure 2.2A-C, SI Tables 4-8). In Experiment 3, larger games began with lower initial accuracy ($\beta = -0.64$, 95% CrI $= [-1.05, -0.25]$) and improved more slowly ($\beta = -0.34$, 95% CrI $= [-0.43, -0.25]$) than smaller games, although group size differences were not reliable in Experiment

1 (SI Table 4), and these experiment 3 differences were not robust in a sensitivity analysis (SI Figure 4C and SI Table 58). Among large groups in Experiment 2, accuracy was higher in the thicker conditions than in the condition with thin interaction structure (SI Tables 5-7), although effects of game thickness were not reliable in Experiment 3 (SI Table 8).

Because each experiment only explored a slice of the full parameter space, we also considered an exploratory analysis that pooled data across experiments, aiming to mitigate the loss in power from running entirely separate regression models. Specifically, we aggregated data from all experiments into a post-hoc mega-analytic model predicting accuracy as a function of repetition block, game thickness (thin v. not-thin) and game size. Overall, we found evidence that accuracy increased over time ($\beta = 0.46$, 95% CrI $= [0.4, 0.52]$) but the rate of increase was reduced for thin games ($\beta = -0.12$, 95% CrI $= [-0.21, -0.02]$) and larger games ($\beta = -0.07$, 95% CrI $= [-0.09, -0.05]$) compared to smaller or thicker games. That is, smaller groups and groups with higher coherence tended to be more accurate, though the magnitude and reliability of these effects varied across individual experiments.

2.2.3 Smaller and higher-coherence groups are more efficient

After establishing that groups were able to communicate accurately, we turned to the challenges faced by describers when deciding how much information to provide. Specifically, we predicted that larger and more heterogeneous groups may initially require more information, but that thicker interaction structure may similarly allow describers to communicate more effectively over time. We tested these predictions using linear mixed-effects models predicting the number of words a describer produced on each trial as a function of condition and block. These models counted all words the describer produced, including after matcher contributions (similar effects were found in models predicting the length of describer's utterances before any matcher contributions, see SI Tables 21-24).

First, as predicted, describers in larger groups produced longer descriptions at the outset than describers in smaller groups (Figure 2.2D-F). This effect held for the continuous manipulation of group size for Experiment 1 ($\beta = 1.6$, 95% CrI =

$[0.62, 2.6]$) as well as the 2-person versus 6-person manipulation in Experiment 3 ($\beta = 7.51, 95\%\text{CrI} = [3.63, 11.3]$). Smaller groups also continued to use shorter descriptions than larger groups over the course of the game. In Experiment 1, the rate at which efficiency increased was similar across different size groups ($\beta = -0.09$, 95% CrI $= [-0.37, 0.18]$). In Experiment 3, larger groups reduced faster than smaller ones ($\beta = -1.22$, 95% CrI $= [-2.06, -0.29]$), but the faster reduction did not fully make up for the longer initial starting point, and was not robust to a sensitivity analysis (SI Figure 4F and SI Table 63).

While thin 6-person games showed a flatter reduction trajectory than thicker 6-person games in Experiment 2 (SI Tables 10-12), there was no reliable effect of game thickness on reduction in Experiment 3 (SI Table 13).

The reduction patterns of description lengths is paralleled by how long matchers took to make selections; across conditions, matchers selected faster in later conditions (SI Figure 9), and the correlation between speed and description length was consistent across experiments (SI Figure 10).

Aggregating across experiments with a mega-analytic model, however, suggested that larger games were associated with steeper reduction ($\beta = -0.36$, 95% CrI $= [-0.51, -0.2]$) from a more verbose starting point ($\beta = 2.12$, 95% CrI $= [1.5, 2.75]$) than smaller games, and thin games had shallower reduction curves ($\beta = 0.79, 95\%\text{CrI} = [0.04, 1.52]$) than thicker games. Overall, then, smaller games used shorter descriptions than larger games across various time points in the experiment, and thinner games reduced less than thicker games.

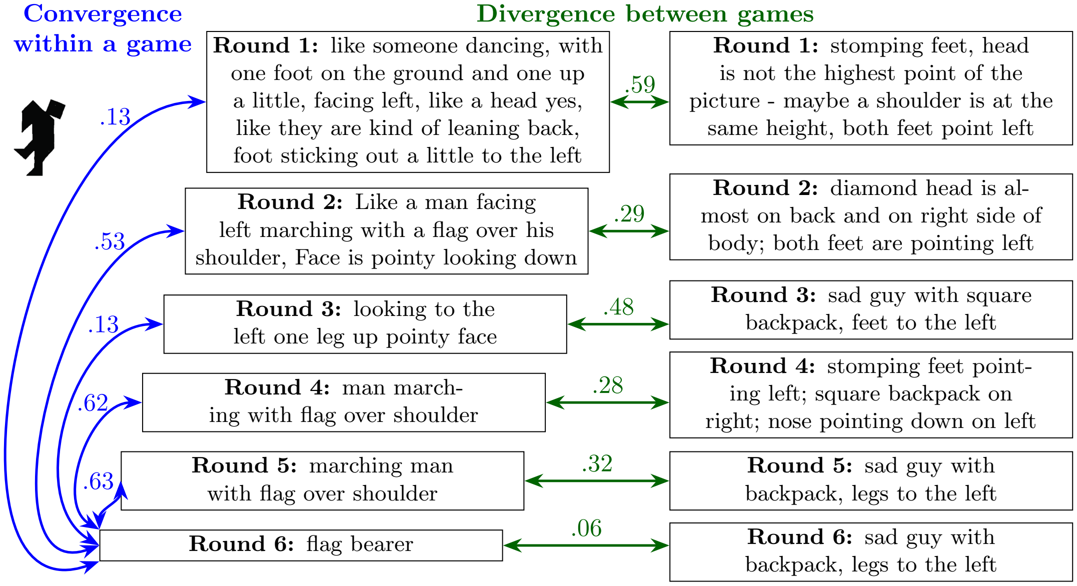
2.2.4 Larger groups make greater use of matcher contributions

As a final measure of group performance, we examined the back-and-forth interactions between the describer and the group of matchers. Matchers use their chat contributions to actively provide feedback, ask questions, offer alternative descriptions, and seek clarification about the describer's referring expressions. Example transcripts from successful games, one in the 6-thick condition and one in the 6-thin condition, are shown in Table **??**. Additional examples are in the SI Tables 1 and 2. Overall, we found that larger groups displayed a higher proportion of trials where at least one matcher produced utterances (SI Figure 6A, $\beta = 0.79$, 95% CrI $= [0.58, 0.98]$), which declined across repetition blocks ($\beta = -0.8$, 95% CrI $= [-0.97, -0.62]$). On an individual level, a matcher in a larger group was more likely to make contributions than a matcher in a smaller group, although each contribution tended to be shorter (SI Figure 7, SI Tables 18, 20). The length of matcher interjections also decreased over time, especially for large groups (SI Figure 6D, $\beta = -0.41$, 95% CrI $= [-0.72, -0.11]$) consistent with the need for early matcher involvement in establishing referential conventions. Emoji use in Experiment 3 followed similar trends (SI Figure 8). Overall, describers in larger groups receive more total input from matchers, suggesting larger groups may require greater participation by matchers to reliably establish common ground.

2.2.5 Descriptions converge faster in groups with thicker channels

In the previous sections, we examined three metrics of communicative performance in groups of different sizes and interaction structures. We confirmed that groups in all conditions replicated the classic patterns of increasing accuracy and decreasing description length. We also found some initial evidence that larger groups may struggle to improve performance in the absence of thick communication channels. Here, we aim to better understand the mechanisms that allow describers to use shorter descriptions without sacrificing accuracy. In particular, we explore the hypothesis that interaction structure and group size affect performance through a *convention formation* process (Clark & Wilkes-Gibbs, 1986). Under a recent model of convention formation (Hawkins et al., 2023), groups are able to leverage their shared history to coordinate on stable expectations about how to refer to particular images. This model makes specific predictions about how interaction structure affects the ability to coordinate, in terms of the available feedback.

First, due to heterogeneity in the group – 6 individuals who may have diverging conceptualizations — a rational describer should provide a strictly more detailed initial description to hedge against multiple possible misunderstandings, as we previously observed. Second, all groups should display the characteristic dynamics of conventions: *stability*, or convergence within group, and *arbitrariness*, or divergence



Round 1: like someone dancing, with one foot on the ground and one up a little, facing left, like a head yes, like they are kind of leaning back, foot sticking out a little to the left

Round 2: Like a man facing left marching with a flag over his shoulder, Face is pointy looking down

Round 3: looking to the left one leg up pointy face

Round 4: man marching with flag over shoulder

Round 5: marching man with flag over shoulder

Round 6: flag bearer

Round 1: stomping feet, head is not the highest point of the picture - maybe a shoulder is at the same height, both feet point left

Round 2: diamond head is almost on back and on right side of body; both feet are pointing left

Round 3: sad guy with square backpack, feet to the left

Round 4: stomping feet pointing left; square backpack on right; nose pointing down on left

Round 5: sad guy with backpack, legs to the left

Round 6: sad guy with backpack, legs to the left

Figure 2.3: Example utterances describing the shown tangram figure produced by two 3-player games in Experiment 1. To measure convergence within a game (blue), we measured the cosine similarity between SBERT embeddings of descriptions and the embedding of the round 6 utterance (taken to be the convention). Higher cosine similarity indicates more similar meaning. To measure divergence between games (green), we measured the similarity between embeddings of utterances from the same round across games.

to multiple equilibria across groups. Third, convergence should be faster when a single individual is consistently in the describer role and when matchers are able to freely respond in natural language, as describers are able to aggregate feedback about the effectiveness of their own utterances from block to block and also immediately correct specific misunderstandings within a given trial.

To assess the dynamics of describer descriptions, we examine the *semantic similarity* of descriptions within and across games. We quantified description similarity by concatenating describer messages together within a trial and embedding this description into a high-dimensional vector space using SBERT. SBERT is a BERT-based sentence embedder designed to map semantically similar sentences to embeddings that are nearby in embedding space. Semantically meaningful comparisons between sentences are made by taking pairwise cosine similarities between the embeddings

(Reimers & Gurevych, 2019).

To measure stability, or convergence within groups, we compared utterances from blocks one through five to the final (block six) description for the same image from the same game. To measure arbitrariness, or divergence across groups depending on group-specific history, we compared utterances produced by different describers for the same image in the corresponding blocks. Figure 2.3 illustrates these two measures with example utterances and their within-game and between-game cosine similarities.

We modeled semantic convergence with a mixed effects linear regression model predicting the similarity between a block 1-5 utterance and the corresponding block 6 utterance as a function of the earlier block number and condition (Figure 2.4A-C; SI Tables 25-29). All conditions showed some convergence toward a conventional "shorthand" for the picture, but the speed of convergence was affected both by group size and channel width. First, we found that smaller groups reached stable descriptions faster than larger games. In Experiment 1, initial similarity was invariant across group size ($\beta = -0.008$, 95% CrI $= [-0.021, 0.005]$), but smaller groups converged faster (Figure 2.4A, $\beta = -0.008$, 95% CrI $= [-0.011, -0.005]$). In Experiment 3, 6-person thick games started off further from their eventual convention than 2-person thick games ($\beta = -0.069$, 95% CrI $= [-0.113, -0.025]$) but closed the gap over time (Figure 2.4C, $\beta = 0.009$, 95% CrI $= [0.001, 0.017]$, this effect was not robust to sensitivity analysis, SI Figure 5C and SI Table 68). Second, thicker games tended to converge faster than thin games (Figure 2.4B-C). In Experiment 3, small thin games started off slightly further from their convention than small thick games, and this gap widened over time ($\beta = -0.025$, 95% CrI $= [-0.033, -0.017]$). Finally, the combination of thin interaction structure and larger group hindered convergence more than either factor individually. Beyond the generally slower convergence in thin games, 6-person thin games showed substantially slower convergence even compared to 2-person thin games in Experiment 3 ($\beta = -0.035$, 95% CrI $= [-0.047, -0.025]$).

Pooling across experiments in a mega-analysis confirms this pattern. Thin games converge less than thick games overall ($\beta = -0.016$, 95% CrI $= [-0.025, -0.008]$), and *large* thin games are especially slow to converge ($\beta = -0.007$, 95% CrI $= [-0.01, -0.004]$). Across games, convergence towards the last utterance was driven by cumulative increasing similarity between pairs of utterances in adjacent blocks (SI

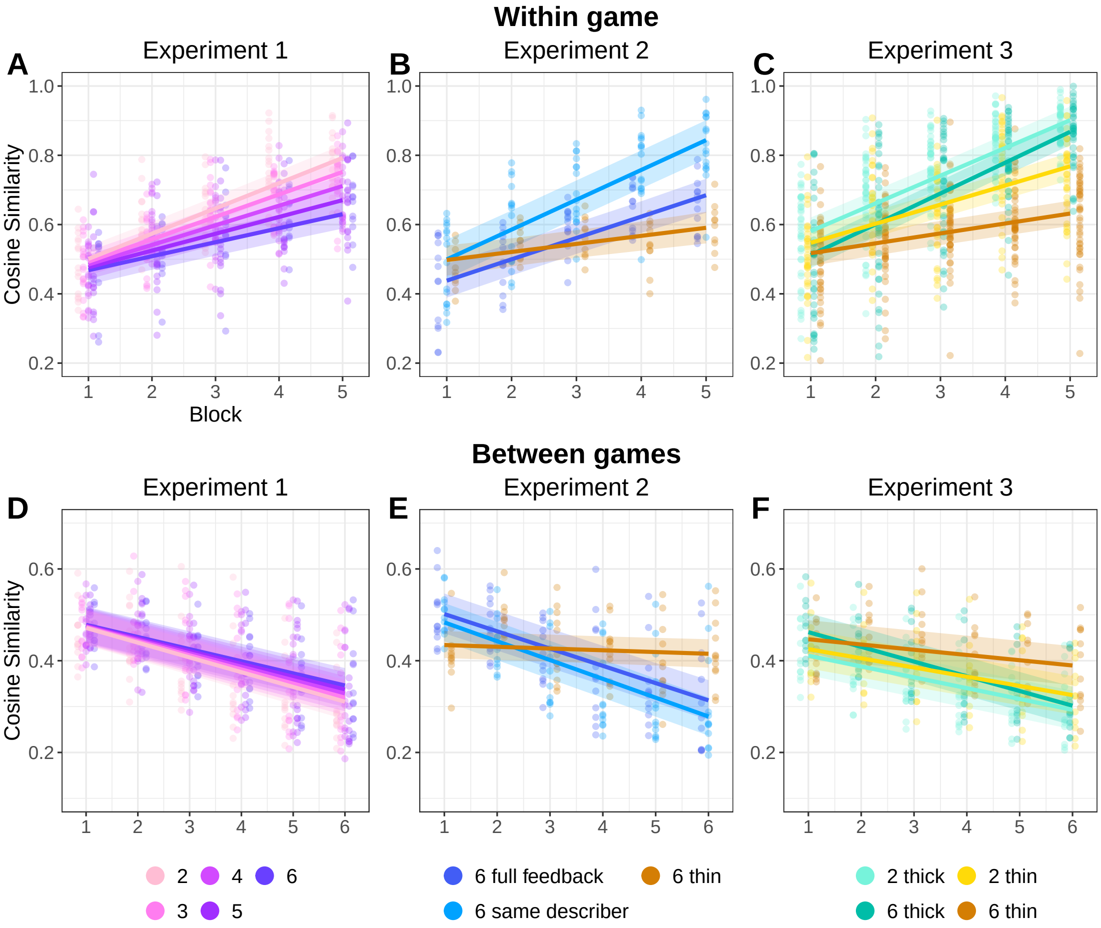


Figure 2.4: Language similarity results measured with pairwise cosine similarity between embeddings of two utterances. (A-C). Convergence of descriptions within games as measured by similarity between an utterance from block 1-5 to the block 6 utterance in the same game for the same image. (D-F). Divergence of descriptions across games as measured by the similarity between two utterances produced for the same image by different groups in the same block. For all, small dots are per game, per block means, and smooth lines are predictions from model fixed effects with 95% credible intervals. Y-axes are truncated, and a few outliers points are not visible.

Figure 12D-F, SI Tables 40-44). In early rounds, descriptions could change substantially between rounds, but by later rounds, many descriptions had already reduced and solidified and varied little round to round. In summary, we found that stable descriptions emerged earlier if the group was smaller, or if the group had a thick interaction structure.

2.2.6 Games with thicker channels diverge from one another more quickly

While groups may initially overlap in their descriptions, including details of shapes or body parts, we predicted that their descriptions would become increasingly dissimilar as groups increasingly adapt to their own idiosyncratic shared history. To test this effect, we constructed a mixed-effects linear regression model predicting the cross-game similarity between a pair of utterances for the same image. A decrease in the similarity between different groups descriptions occurred in every condition, indicating increasing arbitrariness and group-specificity of descriptions (Figure 2.4D-F, SI Tables 30-34). However, different game sizes and interaction structures revealed very different strengths of divergence.

First, smaller games used more group-specific language. In Experiment 1, smaller games diverged more quickly than larger games ($\beta = 0.001$, 95% CrI $= [0.001, 0.002]$). In Experiment 3, 2-person thick games started off more dissimiliar than 6-person thick games, although 6-person games diverged faster and eventually approached the dissimilarity levels of 2-person thick games (SI Table 34). Second, thicker interaction structure was associated with stronger group-specific divergence. In Experiment 3, 2-person thin games diverged more slowly than 2-person thick games ($\beta = 0.004$, 95% CrI $= [0.002, 0.005]$). As with the convergence patterns, large games with thin interaction structures had the flattest trajectories, as thinness and largeness compounded. In Experiment 3, 6-person thin games diverged even less than 2-player thin games (Figure 2.4F, $\beta = 0.017$, 95% CrI $= [0.015, 0.019]$), and in Experiment 2, 6-person thin games barely diverged at all (Figure 2.4E, $\beta = -0.004$, 95% CrI $= [-0.006, -0.001]$). A mega-analytic model confirms this pattern: thin games differentiate less between groups ($\beta = 0.005$, 95% CrI $= [0.004, 0.007]$) and large thin groups differentiate even less ($\beta = 0.004$, 95% CrI $= [0.004, 0.005]$).

As a complement to the embedding analysis, we also examined the frequency of a few classes of words in the descriptions. Literal geometric words (ex. square, triangle, etc) and words for body parts (leg, arm, etc) are common early in games, but decline over repetition in most conditions, to be replaced by more abstract descriptions that do not contain these classes of words (SI Figure 2). The 6-person thin condition, however, retains a higher level of literal geometric and body part words, along with high levels of positional words (above, left, below, etc) and posture words (kicking, standing, seated, etc), with a lower level of utterances that do not contain any of these classes of words.

2.3 Discussion

From classrooms to boardrooms, human communication often takes place in multiparty settings. However, experimental research rarely focuses on such settings, largely due to practical obstacles. In the current work, we asked how convention formation processes, typically studied in dyadic reference games, unfold in larger groups and under varying interaction structures. Across 3 online experiments and 11 experimental conditions, we varied multiple features of interaction structure including group size, modality of matcher contributions, and degree of group coherence. All conditions replicated classic dyadic phenomena: increasing accuracy and efficiency, semantic convergence within games, and differentiation of descriptions between groups. However, we also found that the interaction structure substantially affects how rapidly groups develop partner-specific conventions. Small groups may be able to successfully form conventions under limited feedback, but larger groups require thicker interaction structure. Multi-player groups may therefore reveal key factors which are masked in purely dyadic settings.

Increasing efficiency, for example, has often been taken as an index of group-specific convention formation (Brennan & Clark, 1996; Clark & Wilkes-Gibbs, 1986; S. O. Yoon & Brown-Schmidt, 2014, 2018). In our work, however, we observe distinct patterns for measures of raw utterance length compared to the dynamics of semantic content. In Experiment 3, thin 6-person games showed much less group-specific divergence despite comparable accuracy and efficiency. This gap raises the possibility that it is possible to become more efficient and accurate without negotiating a unified group-specific label. Instead, they may be relying more strongly on the group's priors (Guilbeault et al., 2021). Thus, we encourage measures of semantic content (and not just performance) when evaluating convention formation. The transcripts for these games provide a rich dataset for exploring different ways language is used to form referential conventions.

The causal mechanisms driving group size effects remain unclear. There are many differences between a 2-person group and a 6-person group that could plausibly lead to different outcomes. For example, in a dyad, producers can tailor their utterances to the one matcher, but in large groups, producers must balance the competing needs of different comprehenders (Schober & Clark, 1989; Tolins & Fox Tree, 2016; S. O. Yoon & Brown-Schmidt, 2018). These effects likely vary with the knowledge state and the communication channels available to comprehenders (Fox Tree & Clark, 2013; Horton & Gerrig, 2002a; Horton & Gerrig, 2005). Further work digging into the language used and the interactions between participants might unearth plausible mechanisms for how differences in group size and interaction structure influence outcomes, pointing towards future experimental conditions.

Even within the boundaries of the repeated reference game paradigm, there is a high-dimensional space of possible experiments. We sampled only a few points along a few salient dimensions. In our experiment 3, we grouped some factors together in order to run more games in each condition: a fully factorial design would have been too expensive to power adequately. We instantiated a “thin” channel by limiting matchers to 4 discrete utterances (emojis), but there are other possible restrictions that could be placed on the channel, such as rate-limited typing or explicit time pressure. Future work could explore other dimensions of the interaction structure, introducing pre-existing relationships or familiarity among group members, alternative incentives involving competition and power, or alternative referential targets involving more complex concepts.

A particularly important dimension shaping interaction is the modality of communication, including whether whether the participants use oral or written language, whether they are co-present in the same space, and whether they have visual access to each others’ faces and gestures. Distinct modalities carry distinct affordances and norms. In this work, we relied on a text-based chat modality without allowing co-presence or visual access.

We suspect that the general pattern of effects we see, in terms of group size and coherence, are likely to extend to other modalities. However, different modalities may allow for different strategies that may be more or less sensitive to group size, describer rotation, or different levels of matcher contributions. For instance, in face-to-face oral settings, it may be easier for describers to continuously talk until interrupted, or to monitor the comprehension of individual group members from their facial expressions.

In conclusion, narrowly focusing on the settings that are easy to study in the lab – dyads with rich communication channels – can lead to theories that mispredict how interactions play out in multi-party groups. By studying common ground and coordination across a wider range of interaction structures, we can develop a more nuanced understanding of the obstacles that stand in the way of successful communication and how groups can overcome them. This understanding can inform the design of policies and collaborative platforms that promote effective communication in various contexts, from small-scale conversations to large-scale civic discourse. As remote work and online communication become increasingly prevalent, it is increasingly crucial to understand how the structure of group communication environments shapes the effectiveness of human communication.

2.4 Materials and Methods

Our iterated reference task was implemented with Empirica (Almaatouq et al., 2020), a React-based web development framework for real-time multi-player tasks. Our experiments were designed sequentially and pre-registered individually.[1] We followed the pre-registered analysis plan for each experiment, although accuracy models were not explicitly specified until Experiment 3, and linguistic analyses were only verbally described starting with Experiment 2b. Results from some pre-registered models are omitted from the main text for brevity but are shown in the SI. Exploratory mega-analytic models pooling across the three experiments were not pre-registered.

All materials, data, and analysis code is available at `https://github.com/vboyce/multiparty-tangrams`.

[1] Experiment 1: `https://osf.io/cn9f4` for the 2-4 player groups, and `https://osf.io/rpz67` for the 5-6 player data run later. Experiment 2: same describer at `https://osf.io/f9xyd`, full feedback at `https://osf.io/j5zbm`, and thin at `https://osf.io/k5f4t`. Experiment 3: `https://osf.io/untzy`

2.4.1 Participants

This research was covered by the Stanford IRB under protocol 20009 "Online investigations of language learning". Participants were recruited using the Prolific platform. All participants self-reported as fluent native English speakers on Prolific's demographic prescreen. Experiment 1 took place between May and July 2021, Experiment 2 between March and August 2022, and Experiment 3 in October 2022. Each participant took part in only one experiment and was blocked from participating in subsequent experiments. As games with more participants tended to run longer, we paid participants different rates based on group size, with the goal of a consistent $10 hourly rate. Participants were paid $7 for 2-player games, $8.50 for 3-player games, $10 for 4-player games, and $11 for 5- and 6-player games. When one player occupied the describer role for the entirety of a 6-player game, they were rewarded an additional $2 bonus. Across all games, participants could earn up to $2.88 in performance bonuses.

A total of 1319 people participated across the 3 experiments. We recruited enough participants for 20 games in each condition in experiments 1 and 2 and 40 games per condition in experiment 3. However, due to attrition in filling the games initially and due to participants dropping out of the games, we ended up with fewer games in some conditions. For logistical reasons of matching participants into real-time games, we had to recruit participants in fairly large batches, and so did not have precise control to add new games to replace games that did not fill or had participants drop out early. A breakdown of number of games and participants in each condition is shown in SI Table 3 along with further discussion of recruitment logistics.

2.4.2 Materials

The same 12 tangram images, drawn from Hawkins et al. (2020) and Clark & Wilkes-Gibbs (1986), were used every block. These images were displayed in a 4×3 grid with the order randomized across participants to disincentivize spatial descriptions such as "top left," as the image might be in a different place on the describer's and matchers' screens. To reduce cognitive load from visual search, the locations were fixed for each participant across trials.

2.4.3 Procedure

The experimental procedure was very similar across the three experiments. We first describe the procedure used in Experiment 1 and then describe the differences in later experiments.

Experiment 1 Participants were directed from Prolific to our custom web application, where they were presented with a consent form and a series of instruction pages explaining the protocol. After finishing the instructions, they needed to pass a quiz to proceed. They were then directed to a “waiting room” lobby. Once the lobby filled to the required number of players, the game began. One lobby was filled before another was started; if a participant was waiting for 5 minutes, that lobby timed out, and the participant was paid without completing the experiment. Due to technical constraints with assigning participants to lobbies and games, only games of a single experimental condition could be active at a time. Thus, different conditions were run on different days or times of day.

One of the participants was randomly selected to begin in the role of describer, and the other participants were assigned to the role of matchers. On each trial, the describer saw a fixed array of tangrams with one tangram (privately) highlighted as the *target*. They were given a chat interface to communicate the target to the matchers, who were asked to determine which of the 12 images was the referential target. All participants were free to use the chat box to communicate at any time, but matchers could only make a selection after the describer had sent a message. Once a matcher clicked, they could not change their selection. There was no signal to the describer or other matchers about who had already made a selection. We recorded what all participants said in the chat, as well as who selected which image and how long they took to make their selections.

Once all matchers had made a selection (or a 3-minute timer ran out), participants were given feedback and proceeded to the next trial. Matchers only received *binary* feedback about whether they had chosen correctly or not; that is, matchers who made an incorrect choice were not shown the correct answer (see SI Figure 1 for example feedback). The describer saw which tangram each matcher selected, but matchers did not see one another’s selections. Matchers got 4 points for each correct answer;

the describer got points equal to the average of the matchers' points. These points were translated into performance bonuses at the end of the experiment (1 point = 1 cent bonus). After the describer had described each of the 12 images as targets, in a randomized sequence, the process repeated with the same set of targets, for a total of 6 such repetition blocks (72 trials).

The same person was the describer for an entire block, but participants rotated roles between blocks. Thus, over the course of the 6 blocks, participants were describers 3 times in 2-player games, twice in 3-player games, once or twice in 4 and 5-player games, and once in 6-player games. Rotating the describer was chosen in this first experiment to keep participants more equally engaged (the describer role is more work), and to provide a more robust test of our hypotheses regarding efficiency and convention formation. After the game finished, participants were given a survey asking for optional demographic information and feedback on their experience with the game.

If a participant disconnected from the experiment, the game would stop.

Experiment 2 Experiment 2 consisted of three different variations on Experiment 1, all conducted in 6-player games. Each of these conditions differed from the Experiment 1 baseline in exactly one way. In the *same describer* condition, one person was designated the describer for the entire game, rather than having the describer role rotate. In the *full feedback* condition, all participants were shown what all others had selected as well as the identity of the correct target. This condition was similar to previous dyadic work, such as Hawkins et al. (2020), where the correct answer was indicated during feedback. In the *thin* condition, we altered the chatbox interface for matchers. Instead of a textbox, matchers had 4 buttons, each of which sent a different emoji to the chat. Matchers were given suggested meanings for the 4 emojis during the instruction phase. They could send as many emojis as desired; for instance, they might initially indicate confusion, and later indicate understanding. In addition, for the thin condition, we added notifications that appeared in the chat box marking the time when each player had made a selection.

Experiment 3 The thin channel condition in Experiment 3 was the same as the thin condition in Experiment 2. The thick condition combined the two coherency-enhancing variations from Experiment 2: the same participant remained in the describer role throughout, and full feedback was given about the correct answer and what all other players had selected. Across both conditions in Experiment 3, notifications were sent to the chat to indicate when a participant had made a selection. For experiment 3, game lobbies worked slightly differently, and 5 minutes after the first participant had joined the lobby, the game started if there were at least two participants. Correspondingly, in experiment 3, games did not stop if a player disconnected, instead if there were at least two players still active, the game continued, swapping a player into the role of describer if necessary to continue the game.

2.4.4 Data pre-processing and exclusions

Participants could use the chat box freely, which meant that the chat transcript contained some non-referential language. The first author skimmed the chat transcripts, tagging utterances that did not refer to the current tangram. These were primarily pleasantries ("Hello"), meta-commentary about how well the task was going, and bare confirmations or denials ("ok", "got it", "yes", "no"). We excluded these utterances from our analyses. Note that chat lines sometimes included non-referential words in addition to words referring to the tangrams ("ok, so it looks like a zombie", "yes, the one with legs"); these lines were retained intact.

In Experiments 1 and 2, games did not start if there were not enough participants and ended if any participant disconnected. In Experiment 3, games started after a waiting period even if they were not entirely full and continued even in the event that a participant disconnected (with describer role reassigned if necessary), unless the game dropped below 2 players. The distribution of player counts in games that were initially recruited to be 6 player games is shown in SI Figure 3. The realities of online recruitment and disconnection meant that the number of games varied between conditions. We excluded incomplete blocks from analyses, but included complete blocks from partial games (See SI Table 3). When skimming transcripts to tag non-referential utterances, we noticed that one game in the 6-player thick condition had a describer who did not give any sort of coherent descriptions, even with substantial matcher prompting. We excluded this game from analyses.

2.4.5 Modelling strategy

We fit all regression models in brms (Bürkner, 2018) with weakly regularizing priors. We were unable to fit the full pre-registered mixed effects structure in a reasonable amount of time for some models, so we included the maximal hierarchical effects that were tractable. All model results and priors and formulae are reported in the SI. Models of accuracy used by-group random intercepts only, models of word count used full mixed effect structure, and models of S-BERT similarities used by-group and by-target random intercepts as applicable (see SI Figure 11). Models of matcher accuracy were logistic models with normal(0,1) priors for betas and sd. Models of describer efficiency were run as linear models with an intercept prior of normal(12,20), a beta prior of normal(0,10), an sd prior of normal(0,5) and a random-effect correlation prior of lkj(1). For all of the models of SBERT similarity, we used linear models with the priors normal(.5,.2) for the intercept, normal(0,.1) for betas, and normal(0,.05) for sd. As an additional post-hoc analysis, we ran mega-analytic models combining data across all experiments. For these models, we grouped the 3 thin-ish conditions (2c, and the two thin conditions of experiment 3) as one level, and coded the rest of the conditions as thick-ish. Game size was coded as a continuous measure (2 through 6). The priors for the mega-analytic models were the same as for the per-experiment models described above.

As a sensitivity analysis, we re-ran the primary models on the subset of the data from games that a) completed all 72 trials and b) had the full complement of players the entire time (relevant to 6-player experiment 3 games where games could start or continue with fewer players). Discrepancies are mentioned in the results, and these analyses are depicted in SI Figures 4 and 5 and SI Tables 54-73. We also needed to decide how to handle dropout in Experiment 3, as some of the 6-player games did not retain all 6 players for the entire game. Our decision was to follow an intent-to-treat analysis and treat data as missing completely at random. Note that this choice underestimates differences between 2-player and (genuine) 6-player games by labeling some smaller groups as 6-player groups. We do not know exactly what leads some participants to drop out, but it is possible that some factors may be random (ex.

connection issues) and others may be correlated with performance (ex. frustration because group is struggling).

We do not know whether groups that start and continue at the full size differ from games where some participants drop out. This is potentially an issue across all experiments; in experiments 1 and 2, groups stopped playing if anyone dropped out, and in experiment 3 they kept playing as a smaller group. The number of games in each condition and rates of dropoff are shown in SI Table 3 and SI Figure 3.

Chapter 3

Developmental origins

3.1 Introduction

Learning a language requires learning not only the content of that language, but also how to use the language to communicate. One case study for language use is referential communication, the ability to describe a target so an interlocutor can pick it out from a set of possibilities. Adults show sensitivity to both the visual context and their audience during referential communication, calibrating the description they provide to their beliefs about the interlocutor's knowledge state.

Iterated reference games provide an important paradigm for studying referential communication. In these games, one player repeatedly describes a set of abstract shapes to a partner so they can identify the target images (Boyce et al., 2024; Clark & Wilkes-Gibbs, 1986; Krauss & Weinheimer, 1964; **hawkins2020b?**). Over repetition, features of the initial descriptions are conventionalized as each pair comes to agree on a shared understanding of how to label each image. Success at this task requires mastery of a number of linguistic and communicative skills, including producing adequate initial descriptions, monitoring for comprehension, asking for clarification, and appropriately using the shared conversation history to inform later referring expressions. Studying how children play iterated reference games can provide insight into the developmental trajectory of the ability to produce referential expressions in order to achieve joint understanding.

One influential early study suggested that 4-5-year-old preschoolers struggle with child-child referential communication (Glucksberg et al., 1966). In their paradigm, one child was given a set of 6 blocks in a specific order. Their task was to describe the image on each block so their partner could pick out their corresponding block. As children described and selected blocks, they stacked them on pegs. While 4-5-year-old children succeeded on practice trials with familiar shapes and visual access to each other's blocks, children failed on critical trials where the blocks had abstract drawings and there was no visual access. Even after multiple rounds with the same images, children were not able to correctly order the blocks. Glucksberg et al. (1966) attributed children's communicative failures to their production of ego-centric descriptions that did not account for the other child's perspective.

Similar experiments with older children indicated a gradual improvement through adolescence both for initial accuracy and for the increase in accuracy across repetitions. Still, even the 9th grade sample was noticeably worse than the adult college student sample (Glucksberg & Krauss, 1967; Krauss & Glucksberg, 1969). Given that even teenagers had difficulties with the task, the complex stacking paradigm and the large number of potential targets may have posed task demands that prevented accurate measurement of children's abilities.

More recently, though no evidence has directly contradicted Glucksberg et al. (1966), a number of other studies have revealed early emerging skills in the preschool years that support the use of referential communication. Preschool-aged children are faster to select a target when it is referred to in a consistent way (Graham et al., 2014; Matthews et al., 2010) and 6-year-olds are more likely to use consistent referential expressions when their partner is consistent (Köymen et al., 2014), suggesting that young children are sensitive to the norm of consistent descriptions in cooperative communication. Further, preschool-aged children adapt the informativeness of their referential expressions based on the visual content available to their interlocutor (Matthews et al., 2006; Nadig & Sedivy, 2002; Nilsen & Graham, 2009). By age 5, children integrate information about other's perspectives into their comprehension of utterances (San Juan et al., 2015). Overall, by late preschool, children can can reason about others' perspectives to communicate more effectively.

While children are sensitive to others' perspectives, they still struggle to appropriately tailor the specificity of their utterances to the visual context, often resulting

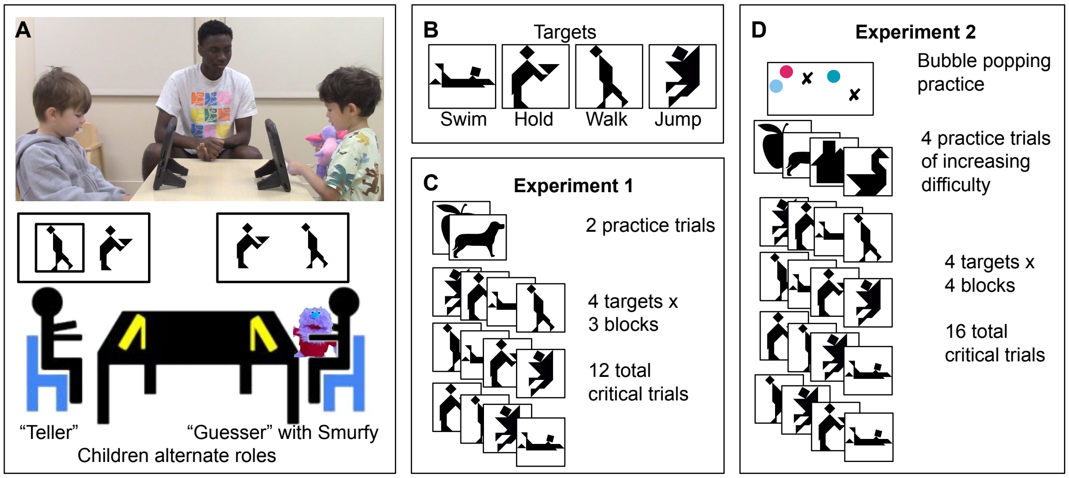


Figure 3.1: Experimental setup and procedure. Panel A shows the experimental setup. Panel B shows the 4 possible targets; names for targets are for cross-reference with later figures only. Panel C shows the procedure for Experiment 1; within critical blocks targets were ordered randomly. Panel D shows the procedure for Experiment 2.

in under-informative utterances (Leung et al., 2024; Matthews et al., 2012). When children must describe one of two very similar images, 4 and 5-year-olds sometimes neglect to mention the relevant features, although they do better when playing in an interactive game than when not (Grigoroglou & Papafragou, 2019). By 5 years old, children are sensitive to over- and under-informative utterances, asking for clarification on some under-informative utterances and taking longer to make selections on over-informative utterances (Morisseau et al., 2013).

A common thread among many developmental studies is that children perform better when the cognitive load of the task is reduced, i.e. when there are fewer possible referents (Abbot-Smith et al., 2016; San Juan et al., 2015). These findings paired with evidence of emerging communicative skills in young children suggest that tailored referential expression production is possible in children, but it is likely to be masked when the task demands are too high.

Since Glucksberg et al. (1966), few studies have revisited the question of whether children can successfully communicate in an iterated reference game. In one study, 8-10-year-olds exhibited adult-like patterns of increasing accuracy, increasing speed, and shorter descriptions across repetitions, but children's accuracy was still far below adult performance and highly variable between dyads (H. P. Branigan et al., 2016). 4-6-year-olds successfully used conventionalized gestures with their partners in an iterated reference game where children could only use gestures to communicate (Bohn et al., 2019). 4-8-year-old children succeeded at playing an iterated reference game with a parent using a simple, child-friendly tablet-based task (Leung et al., 2024). In this task, even 4-year-olds had an initial accuracy above 80%, which rose to above 90% in later repetitions (Leung et al., 2024). Together, these findings provide evidence of young children's ability to communicate about novel referents under the right conditions.

Given the task demands in Glucksberg et al. (1966) and work showing that children can succeed in a less demanding paradigm, here we revisit the question of child-child referential communication. In the present study, we re-examine young children's ability to establish effective referring expressions with each other in an iterated reference game using a simplified tablet-based paradigm. Across an initial study and a replication including a total of 51 pairs of 4-5-year-old children, we found that preschool-aged children were successful in an iterated reference game, suggesting that children's capacity to construct effective referring expressions in novel contexts emerges earlier than once claimed.

3.2 Experiment 1

3.2.1 Methods

Our goal for Experiment 1 was to test young children's ability to coordinate to produce descriptions of abstract shapes that their partner could understand. Young children can be very sensitive to task demands and cognitive load (Carruthers, 2013; Keen, 2003; Turan-Küçük & Kibbe, 2024), so we adapted the experimental framework from Leung et al. (2024), and further simplified it by reducing the total pool of targets and the number of trials. This experiment was pre-registered at anonymized link.

Participants

4 and 5-year-old children were recruited from a university preschool during the school day. Children played with another child from the same class. Experiment 1 was conducted between June and August 2023. Pairs of children were included in analyses if they completed at least 8 of the 12 critical trials. We had 19 games that completed 12 critical trials, and 1 game that completed 11 critical trials. Of the 40 children, 21 were girls, and the median age was 57 months, with a range of 48-70 months.

Materials

For the target stimuli, we used four of the ten tangram images from Leung et al. (2024), chosen based on visual dissimilarity (Figure 4.1B). We coded the matching game using Empirica (Almaatouq et al., 2020), hosted it on a server, and then accessed the game on tablets that were locked in a kiosk mode so children could not navigate away from the game.

Procedure

Once a pair of children agreed to play the game, a research assistant took them to a quiet testing room. Children were introduced to a stuffed animal "Smurfy" who wanted to play a matching game. Children sat across a table from each other, each with a tablet in front of them (Figure 4.1A). On each trial, one child was the "teller" and saw a black box around one of two images on their screen and was asked to "tell Smurfy what they see" in the black box. The "guesser" saw the same two images in a randomized order and tried to select the described image to help Smurfy make a match. When the guesser selected an image, both children received feedback in form of a smiley or frowny face and an excited or disappointed sound. After each trial, children switched roles. Children passed Smurfy back and forth to keep track of whether they were the "guesser" or "teller" on a given trial.

Children completed two warm-up trials with black and white images of familiar shapes, followed by 3 blocks of the 4 target images (Figure 4.1C). Targets were randomly paired with another of the critical images as the foil.

The experimenters running the game did not volunteer descriptions, but they did scaffold the interaction, prompting children to describe the images, and sometimes repeating children's statements (especially when utterances were inaudible or the child did not respond immediately; this aspect of the procedure was modified in Experiment 2). The entire interaction was video-recorded.

Data processing

Children's selections and the time to selection were recorded from the experiment software. Children's descriptions were automatically transcribed from the video using Whisper (Radford et al., 2022) for the first pass and then hand-corrected by experimenters. Transcripts were hand-annotated for when each trial started, who said each line, and what referential descriptions were used. We excluded trials where the "teller" did not produce a description, or where all description was unintelligible and impossible to transcribe. After exclusions, we had 231 trials remaining.

Statistical analyses were run in brms (Bürkner, 2018) with weakly informative priors. We report estimates and 95% credible intervals. The experimental set-up, analysis code, de-identified transcripts, and performance data for both experiments is available at anonymized repo.

3.2.2 Results

Accuracy and speed

Our primary measure of interest was whether children could accurately communicate the intended target. To test for changes in accuracy over time, we fit a Bayesian mixed effects logistic regression predicting accuracy.[1] Children's accuracy was above chance (Odds Ratio: 3.00 [1.14, 8.09]), and their accuracy slightly increased over the game (OR of one trial later: 1.17 [1.03, 1.39], Figure 3.2). This level of accuracy is generally in-line with accuracies from 4-year-olds playing with their parents in Leung et al. (2024) and indicates that children can understand and succeed at the task.

As another measure of children's performance, we looked at how long children spent on each trial. We ran a Bayesian mixed effects linear regression predicting the

[1] correct.num∼ trial.num + (trial.num|game) + (1|target)

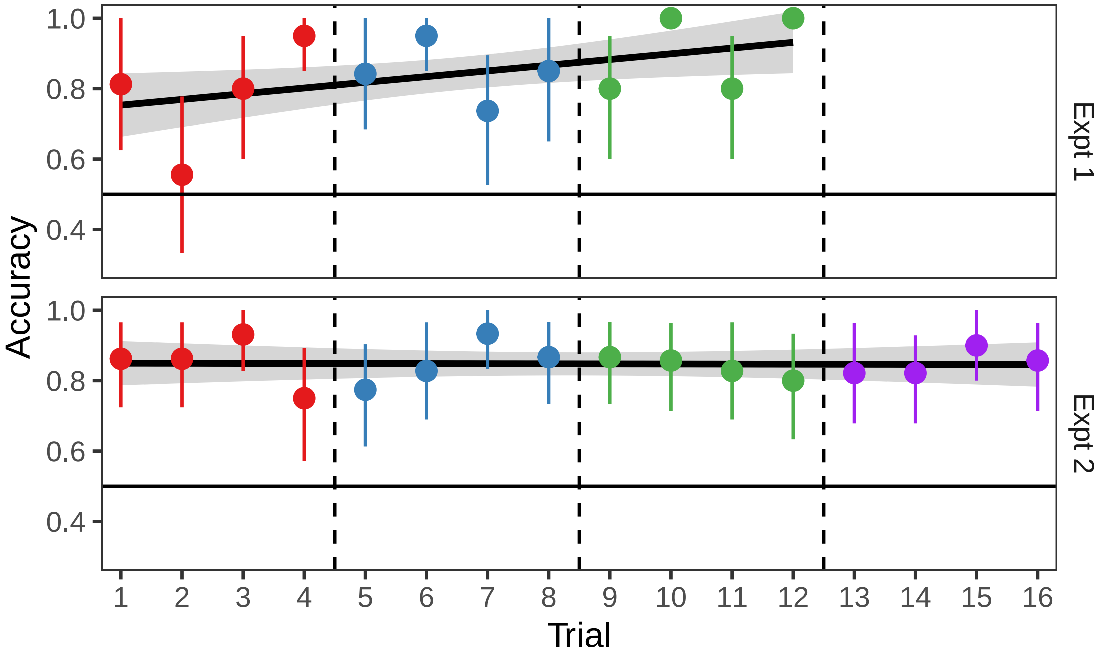


Figure 3.2: Children's accuracy at selecting the correct target over time. Error bars are bootstrapped 95% CIs with a linear trend line overlaid.

time to selection in seconds.[2] The first critical trial averaged 27.48 [20.48, 34.53] seconds, and children got faster over time (-1.22 [-1.90, -0.53] seconds / trial). Children were able to achieve the same accuracy in less time, suggesting that they were becoming more efficient at completing the task.

Description length

In iterated reference games with adults, description lengths usually shorten over repeated references (Boyce et al., 2024; Clark & Wilkes-Gibbs, 1986; **hawkins2020b?**). We were curious if children's descriptions would display the same trend, so we ran a Bayesian mixed effects linear regression predicting the number of words in the description the "teller" produced.[3] On the first critical trial, descriptions averaged 3.66 [2.56, 4.74] words, and description length was relatively stable over time (change of 0.12 [-0.04, 0.28] words per trial, Figure 3.3). Thus, children's increasing speed was

[2] time.sec∼ trial.num + (trial.num|game) + (1|target)

[3] words∼ trial.num + (trial.num|game) + (1|target)

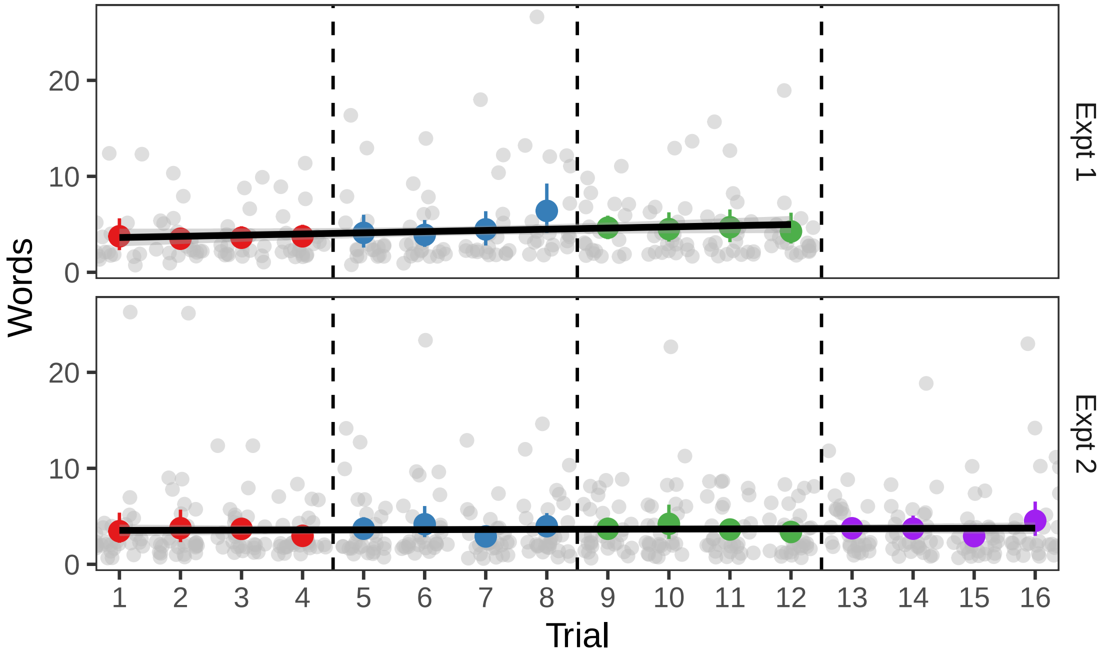


Figure 3.3: Length of description produced by the teller each trial. Grey dots are individual data points, colored dots are per trial means with bootstrapped 95% CIs.

not from shorter utterances, but instead from some combination of improved task understanding, faster utterance planning, and faster decisions of what to select.

Some examples to illustrate the variety of effective descriptions children employed are shown in Table 3.1 (these specific examples are from Experiment 2, but both experiments had similar distributions of descriptions).

Convergence

While description length is often used as a proxy for measuring convention formation, it does not capture semantic overlap between utterances. Boyce et al. (2024) introduced a more sensitive measure of semantic convergence that compares the content of utterances using word embeddings to trace how similarities within and across games change over time. With only 3 blocks of descriptions, we do not expect semantic similarity for descriptions of a given target to show any meaningful change over time. However, we can test for a more coarse measure of sensitivity to partner: whether

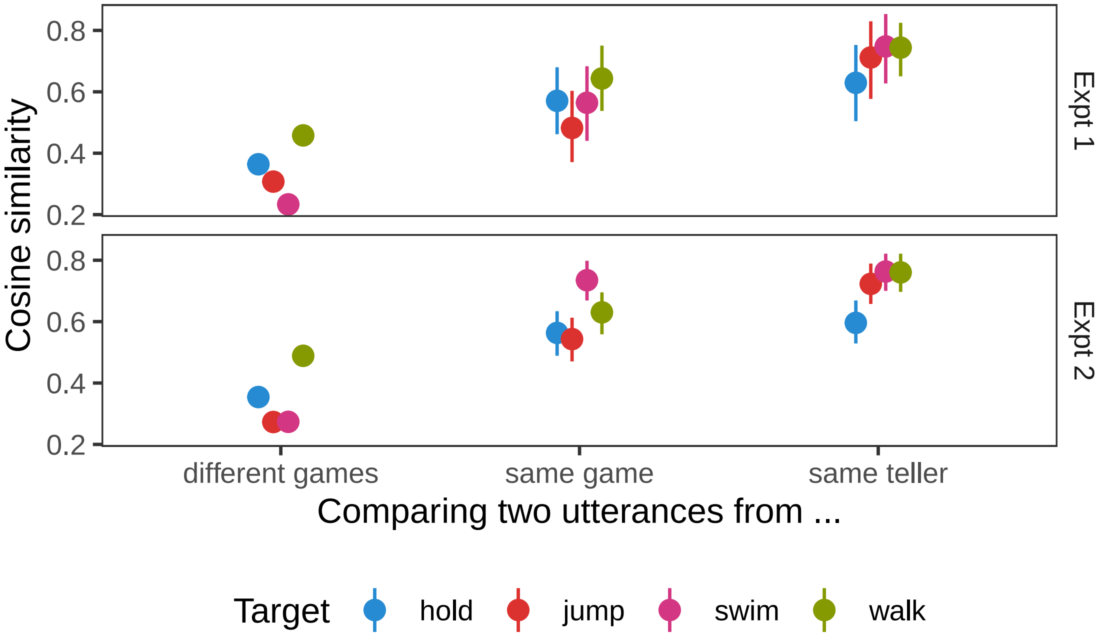


Figure 3.4: Semantic similarity between pairs of descriptions from different sources. Dots are means and lines are bootstrapped 95% CIs.

children's utterances are more like their partner's than children in other games'. If children are fully ego-centric (as suggested by Glucksberg et al., 1966), their choices of descriptions would be independent from their partners.

Following the methods of Boyce et al. (2024), we embedded each description in a semantic vector space using S-BERT (Reimers & Gurevych, 2019), and then used the cosine between embeddings as a measure of semantic similarity.

We compared the semantic similarities between descriptions of the same target based on who produced the description (Figure 3.4). We used a Bayesian mixed effects linear regression to predict similarity.[4] Utterances were more similar if they came from the same partnership (0.222 [0.184, 0.260]). Utterances were slightly more similar still if they came from the same person (0.135 [0.077, 0.193]), which is expected since children are likely to be fairly consistent with themselves. However, children used descriptions that were much more similar to their partner's than to other children's (Figure 3.4), indicating sensitivity to their partner's expressions.

[4] sim∼ same_game + same_speaker + (1|target)

3.2.3 Discussion

In Experiment 1, we adapted the paradigm of Leung et al. (2024) for pairs of children, taking an already simple set-up and making it shorter. Our goal was to see if young children were at all able to provide adequate descriptions, so children received a lot of scaffolding around the experimental interaction. Sometimes, this scaffolding included experimenters echoing children's descriptions, which could potentially influence children's responses. In Experiment 2, we repeated the same paradigm, with a tighter experimental script and a larger sample size.

3.3 Experiment 2

3.3.1 Methods

As Experiment 2 was very similar to Experiment 1, we focus on the changes made compared to Experiment 1. Experiment 2 was pre-registered at anonymized link.

The biggest change between the experiments was increasing the number of repetitions of target stimuli from 3 to 4 (from 12 to 16 trials). The greater number of trials in Experiment 2 made it possible to look for changes over time that could be indicative of convergence to shared descriptions within a game and divergence between games.

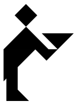
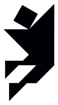
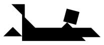
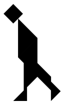
Participants

Experiment 2 was run between March and August of 2024, at the same preschool as Experiment 1. No children participated in both experiments. 30 pairs of children completed all 16 critical trials, and 1 pair of children completed 10 critical trials. Our target age range was 4 and 5-year-olds, but one older 3-year-old was unintentionally included. Of the 62 children, 30 were girls, and the children had a median age of 56 months, and a range of 45-69 months.

Materials

The same 4 critical images were used as in Experiment 1. In response to some children struggling with the abrupt switch from familiar to non-nameable shapes,

Table 3.1: Example descriptions children successfully used to identify different target images in Experiment 2.

● person ● a person holding a sandwich ● a people carrying a box of dirt ● a monster ● someone holding a plate and giving it to a restaurant and has watermelon	
● vampire ● hopping ● a person flying ● a person ● a kite ● a triangle with a head on it with feet ● somebody skydiving, not in the airplane	
● racecar ● airplane ● alligator ● a person fell down ● a boat ● a person that's in a race car that has one triangle and two triangles	
● person ● a person walking ● a person looking down ● a people, but it doesn't have any arms	

we introduced more practice trials for Experiment 2. We used a total of 4 practice trials to provide a gradient from familiar shapes to less recognizable, blockier shapes (Figure 4.1D).

Procedure

The procedure was much the same as Experiment 1(Figure 4.1D). We added an initial “bubble popping” exercise to give children practice tapping the tablet appropriately (this was an issue for some children in Experiment 1). The experimental script was fully written out and memorized by experimenters so children all received the same instructions. We wrote up contingency statements that the experimenter could use to prompt children who were not giving descriptions or making selections. Experimenters helped with game mechanics such as whose turn it was to tell and who should press the screen, but avoided contributing or repeating any content about the images or the descriptions.

Data processing

Data were processed in the same way as Experiment 1. After excluding trials where children did not give a description or where the experimenter echoed a child’s description, we had 466 trials total.

3.3.2 Results

Accuracy and speed

In Experiment 2, children’s accuracy was above chance (Odds Ratio: 5.95 [3.07, 11.89]) and relatively stable over time (OR of one trial later: 1.01 [0.94, 1.09], Figure 3.2). The first critical trial averaged 22.06 [15.86, 28.58] seconds, and children got faster over time (-0.70 [-0.99, -0.41] seconds / trial). Children were initially faster in Experiment 2 than Experiment 1, possibly due to the increased number of practice trials and pre-training on how to press the screens. Taken together, we find more evidence that children can successfully communicate with each other about these abstract shapes, and do so with increasing efficiency.

Description length

The average length of descriptions on the first trial was 3.44 [2.23, 4.75] words and description length was relatively stable over time (change of 0.02 [-0.05, 0.09] words / trial, Figure 3.3). This finding is comparable to Experiment 1, again finding that children produce short utterances without much change in length over time.

Convergence

As a coarse measure of partner-sensitivity, we repeated the semantic analysis from Experiment 1. Utterances were more similar if they came from the same partnership (0.270 [0.243, 0.297]) and were slightly more similar if they came from the same person (0.097 [0.059, 0.132]).

As Experiment 2 had 4 blocks, we examined whether descriptions were converging semantically toward the final description. We compared the utterances from the first three blocks to the descriptions in the last block using a Bayesian mixed effects linear regression predicting similarity.[5] Over the first three blocks, descriptions became increasingly similar to the last block description (0.042 [0.007, 0.078]). Descriptions were more similar if they came from the same child, which is expected as a sign of internal consistency (0.067 [0.006, 0.127]). Although over time descriptions did get more similar to the last block utterance, the semantic distance between adjacent block utterances was relatively constant (0.009 [-0.026, 0.044]).

Partnerships often diverged from one another as groups focused on distinct aspects of the image. We tested whether descriptions in different games diverged over time using a Bayesian mixed effects linear regression.[6] As the games progressed, descriptions to the same target from different games became slightly less similar (-0.013 [-0.018, -0.008]), which indicates that games are converging to different conventions. These patterns of increasing similarity within games and increasing divergence between games qualitatively match the patterns found for adults (Boyce et al., 2024).

[5] sim∼ earlier_block.num + same_speaker + (1|game1) + (1|target)

[6] sim∼ block.num + (1|target)

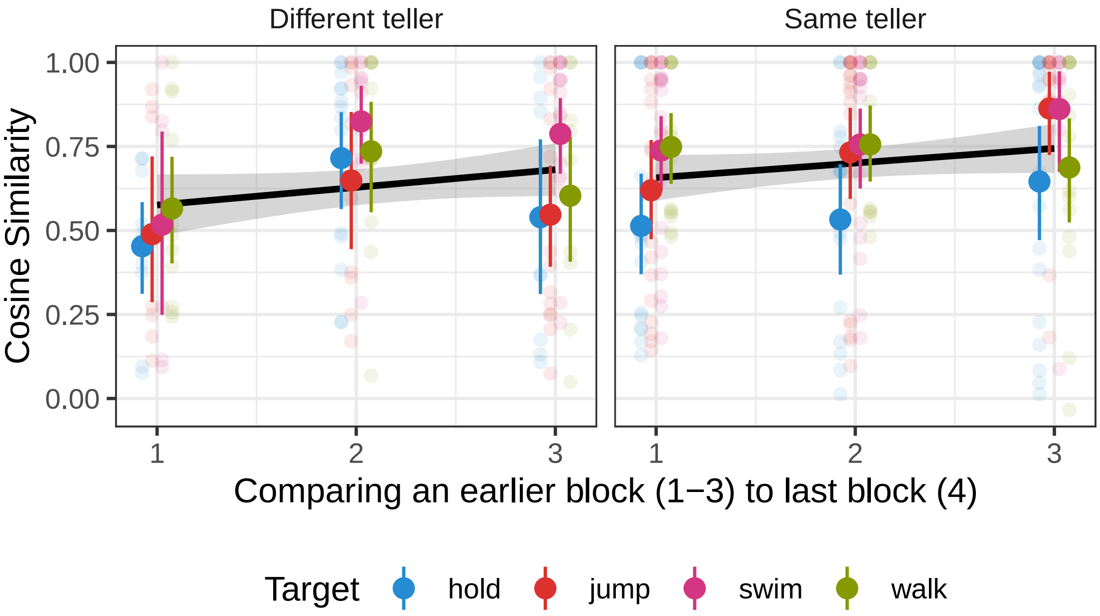


Figure 3.5: Semantic similarity between descriptions from earlier blocks (1-3) and the last block in Experiment 2. Heavy dots are means with bootstrapped 95% CIs; light dots are individual values.

3.4 Joint analysis

As the two experiments were similar to one another, we re-ran models pooling the data across the two experiments, using experiment number as a random effect. Pooling the two experiments, children's accuracy was above chance (OR: 4.64 [1.42, 12.16]) and accuracy numerically increased over the course of the game, although the credible interval included 0 (OR of one trial later: 1.04 [0.98, 1.12]). Descriptions produced by tellers averaged 3.62 [0.09, 7.38] words on the first trial, and description length was relatively stable over time (change of 0.04 [-0.02, 0.11] words per trial). Utterances were more similar if they came from the same partnership (increase in cosine similarity: 0.257 [0.234, 0.279]) and were slightly more similar still if they came from the same person within the partnership (0.109 [0.078, 0.140]).

We might expect that whether a description was successful influences whether the same description, or a variant of it, is employed in future rounds. Intuitively, successful descriptions can be copied and built upon, while unsuccessful descriptions should be replaced by a fresh attempt. To test whether accuracy is predictive of similarity to future descriptions, we ran a post-hoc Bayesian linear model predicting similarity to the next block description in terms of accuracy.[7] Descriptions that elicited a correct response were more similar to the next block description (0.145 [0.055, 0.236]) with no substantial interaction with block number or whether both descriptions came from the same teller. This pattern of results is consistent with the expectation that children are more likely to stick to their own description or repeat the other child's description if it was previously successful.

3.5 General discussion

Prominent early studies claimed that young children cannot overcome their egocentrism to coordinate with each other in reference games (Glucksberg et al., 1966). However, more recent developmental work has found that young children show emerging communicative and pragmatic sensitivity, especially when cognitive demands are low. Here, we revisited the question of preschoolers' performance in child-child reference games, using a scaffolded paradigm to reduce extraneous task demands.

Across 2 experiments and 51 pairs of 4 and 5-year-old children, we tested how well children could produce referential expressions that allowed their partner to find a matching abstract shape. Children varied substantially in what sorts of descriptions they produced, but overall accuracy was high (85%), indicating that children were generally able to produce adequate descriptions. Additionally, children's utterances showed signs of converging toward conceptual pacts. While this task is substantially scaled down relative to measures used for adult competence, it does suggest that the relevant communication skills are present at least in rudimentary form by the end of the preschool years.

Unlike adults, children did not display an increase in accuracy or a shortening of referential expressions over the course of the game. Still, our findings show that descriptions became increasingly similar to descriptions in the last block and that successful utterances were more similar to future utterances, suggesting that children are adapting their descriptions as the game unfolds. These null findings are likely a result of initially high accuracy in the first block and initially short utterances that

[7] sim$\sim$ earlier_block.num $\times$ correct + same_speaker $\times$ correct + (1|game1) + (1|target) + (1|expt)

leave little room for reduction.

It is unclear to what extent the uniformly short descriptions we observed are a product of the simplified task or children's behavioral differences from adults. In this case, the low number of options and relatively easy-to-describe shapes may have obviated the need for long initial descriptions. Indeed, adult controls in Leung et al. (2024) used shorter initial descriptions than adults in studies with larger arrays of harder to distinguish images (Boyce et al., 2024; **hawkins2020b?**). However, young children may also struggle to produce longer descriptions, and young children may be more willing to take guesses when adults would seek additional clarification. Especially in light of other work suggesting that conceptual pact formation and reduction in utterance length sometimes decouple in adults (Boyce et al., 2024), further empirical work on the factors driving verbosity in reference games is warranted.

The generalizability of our results is limited by the target population, the target images, and the task structure. We sampled a convenience population of children at a university nursery school. The set of tangram images may be easier to distinguish and have higher codability than other target images used in adult reference games. We specifically targeted children's abilities to construct referring expressions that can be jointly understood, so children were provided scaffolding around taking turns and talking to their partner. Thus, children's performance should be taken as a proof-of-concept about ability, rather than a claim about how generally children spontaneously demonstrate these abilities.

In the broader picture of language acquisition, there is debate over the timing of the emergence of communicative and pragmatic abilities relative to the acquisition of grammar and meaning. On one side, children seem to learn literal semantics far before they display an understanding of some pragmatic implicatures (Huang & Snedeker, 2009; Noveck, 2001); on the other, sensitivity to communicative intent is an early emerging skill that develops in parallel with linguistic knowledge and may bootstrap language learning (Bates, 1974; Bohn & Frank, 2019; Tomasello, 2008). Our current findings are most consistent with a gradual development of children's communicative and linguistic skills, where the skills emerge early and then are refined over time, as children's cognitive capacities increase. At 4-5 years old, children are already able to establish novel referential conventions with one another as part of their broader ability to communicate and coordinate.

Chapter 4

Opacity and processing

4.1 Introduction

When a teen says about someone that "he's got rizz," what does this mean? The idea that teen slang is arbitrary and opaque to outsiders (i.e., older generations) is enough of a cultural touchstone that late night comedy shows have segments about it. Teens are far from the only ones to have in-group naming conventions; many communities form stable linguistic conventions including professional jargon, regionalisms, and of course slang.

The formation of linguistic conventions between individuals is often studied experimentally in iterated reference games. In these games, a describer tells their partner how to sort or match a series of abstract images (e.g., Clark & Wilkes-Gibbs, 1986; **hawkins2020b?**). Over repeated rounds of referring to the same targets, pairs develop conventionalized nicknames for the target images. These nicknames are often partner-specific, in that different pairs develop different nicknames for the same targets. When describing the targets to a new person, describers return to more elaborated descriptions, indicating an expectation that prior conventions are not appropriate descriptions to use for new matchers (Hawkins, Franke, et al., 2021; Wilkes-Gibbs & Clark, 1992; S. O. Yoon & Brown-Schmidt, 2018).

For one-shot reference games, the choice of referring expression can be described as a process of recursive inference, in which speakers reason about how listeners will interpret their utterance and vice versa. This process is described in Bayesian pragmatics models such as the Rational Speech Acts model (RSA) (Frank & Goodman, 2012; Goodman & Frank, 2016). The Continual Hierarchical Adaptation through Inference model (CHAI) builds on RSA by adding rules for how agents update their belief distributions after each interaction, to account for the dynamics of repeated interaction in iterated reference games (Hawkins, Liu, et al., 2021). A key factor permitting variation in referring expressions is lexical uncertainty; RSA models incorporating lexical uncertainty (Bergen et al., 2016; Potts et al., 2015) treat listeners as Bayesian agents who jointly infer the meaning of a speaker's utterance and their model of the speaker's lexicon. Models incorporating lexical uncertainty have been used to model children's word learning (Bohn et al., 2022) as well as person- and group-level differences (Hawkins, Liu, et al., 2021; Schuster & Degen, 2020). Usually the scope of person-to-person variation in lexica is constrained, so it is more akin to learning a hierarchical set of tweaks to a lexicon so some words have slightly different extensions for different speakers and different circumstances, rather than learning completely arbitrary meanings for each person (Hawkins, Liu, et al., 2021; Schuster & Degen, 2020; but see Misyak et al., 2016).

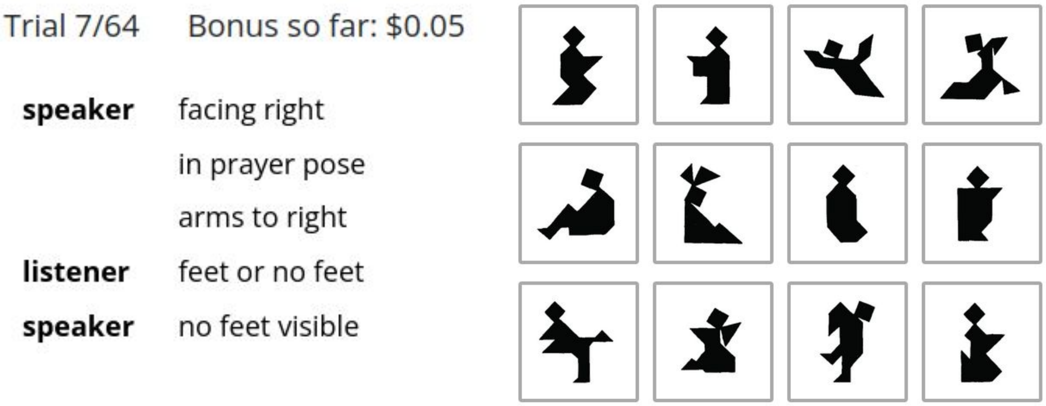
Conventions in iterated reference games are formed between people without a salient shared group identity prior to the interaction, but people do not use the conventions with new partners. How opaque are the temporary linguistic conventions created in reference games: are they opaque like "rizz" or interpretable like "roundabout"? From a theoretical perspective, measuring this opacity can help inform the development of models like CHAI that aim to account for lexical change in conversation. We attempt to answer this question here.

One way to measure the opacity of a referring expression is to look at the semantic distance between the signifier and the referent. Expressions that are more transparent are those where signifiers and referents are semantically close, such that anyone with the same general lexicon can identify the appropriate referent given the signifier. In contrast, expressions that are opaque have signifiers and referents that are semantically distant in the lexicon, such that the relations are arbitrary and inaccessible without additional clues such as the partner-specific conversation history.

One option for measuring the transparency of referring expressions is to use vision–language models to operationalize a shared semantic space for both language and images. Computational methods have enabled the embedding of various stimuli (including images and text) into high-dimensional feature spaces; these embeddings have properties which suggest that they are reasonable approximations of humans' semantic spaces, including similarity in representational geometries (e.g., Grand et al., 2022; Muttenthaler & Hebart, 2021). Indeed, embeddings from neural network models have been used as a form of semantics in a range of reference game scenarios (e.g., Gul & Artzi, 2024; Ji et al., 2022; Kang et al., 2020; Le et al., 2022; Ohmer et al., 2022). In particular, such embeddings can be treated as the default context- and speaker-independent lexicon, since they are not updated to account for convention formation within an iterated reference game.

A second option for measuring the transparency of referring expressions is to measure how often naïve humans, who were not part of the group who formed the convention, can correctly associate the target referent with the referring expression. Prior work with naïve matchers has been limited and has focused on the role of conversational history. In this work, naïve matchers tend to do better the more their observation history resembles that of the original conversation—when hearing descriptions in order instead of in reverse order (Murfitt & McAllister, 2001), when listening to the entire game instead of starting in the third round (Schober & Clark, 1989), and when seeing yoked trials from a single game rather than trials sampled across 10 games (**hawkins2023a?**). In these studies, naïve matchers had worse accuracy than in-game matchers, but their performance was still far above chance, suggesting that the convention–target relationship is not purely arbitrary. In fact, even when pairs of participants try to obfuscate their meaning, overhearers can still do quite well at identifying the target referents (Clark & Schaefer, 1987). Nonetheless, receiving more context from an interaction—and in particular having that context be in order—is beneficial to matchers.

In the current work, we measure the opacity of the referring expressions created in iterated reference games. We use conversations from Boyce et al. (2024), who created a large corpus of iterated reference games that were played online using a chatbox for communication. This corpus is made up of 6-round iterated reference games using the



Trial 7/64 Bonus so far: $0.05

speaker facing right
in prayer pose
arms to right

listener feet or no feet

speaker no feet visible

Figure 4.1: Experimental setup. Naïve matchers read transcripts from trials in reference games from Boyce et al. (2024) and selected which image they thought was being described. Matchers recieved bonus payments for correct selections.

same 12 target images. Games varied in how large the describer–matcher groups were (2–6 participants) and how "thick" the communication channels were (for example, if matchers could send back messages or just emoji). The varied conditions within a consistent framework allowed us to test how the opacity of referring expressions varies depending on the conditions the referring expressions came from.

We use both human experiments and models to assess when and why expressions are opaque or understandable to outside observers. We first present a computational approach using a vision–language model to measure the semantic similarities between referring expressions and their targets, and we validate our model against naïve human matchers (Experiment 1). We then use both naïve human matchers and the model to compare the opacity of referring expressions across different game conditions and time points (Experiment 2). Finally, we address the role of conversation history by comparing naïve matcher performance on game transcripts presented in order versus out of order (Experiment 3).

4.2 Task setup

4.2.1 Materials

We drew our referring expressions from Boyce et al. (2024), excluding utterances that were marked as not containing referential content. For our naïve matcher experiments, we sampled different subsets of this corpus. Within the subsets, we excluded transcripts that contained swear words or crude or sexual language. For the computational model, we used the entire corpus, and pre-processed the text by concatenating all the referential messages sent by the describer for a given trial.

4.2.2 Experimental procedure

We recruited English-speaking participants from Prolific. On each trial, participants saw the full transcript from that trial, containing all the chat messages marked by whether they were from the speaker or a listener. Participants selected the image they thought was the target from the tableau of 12 (Figure 4.1). Participants received feedback on whether they were right or wrong on each trial. Except when the specific viewing order was part of the experimental manipulation, we randomized the order of trials, subject to the constraint that the same target could not repeat on adjacent trials. The task was implemented in jsPsych (Leeuw et al., 2023). We paid participants $10 an hour plus a bonus of 5 cents per correct response. All our experimental code is at this anonymized repo.

4.2.3 Computational models

We used the Contrastive Language-Image Pretraining model (CLIP; `clip-vit-large-patch14`) as a comprehender model for our domain (Radford et al., 2021). CLIP is a vision-language model that uses a text transformer and a vision transformer to embed text and images into the same space, trained to maximize the similarity between representations of images and their English captions. It is a natural choice for reference games, as the model is trained to estimate the correspondence between images and phrases in natural language. We ran CLIP for the concatenated describer utterances and all 12 tangram shapes. For each utterance, we computed probabilities for each

Table 4.1: Cross-validated accuracies for classifiers. Standard deviations in accuracy across the 10 folds are shown in parentheses. Best performance within each model class is underlined, and best overall performance is bolded.

Classifier	Accuracy
Random baseline	0.08
CLIP without readout	0.31
Logistic regression	
No penalty	0.50 (0.01)
L2 penalty	0.50 (0.01)
Random forest	
10 estimators	0.46 (0.02)
50 estimators	0.51 (0.02)
100 estimators	0.52 (0.02)
500 estimators	0.52 (0.02)
Gradient-boosted tree	
10 estimators	0.48 (0.02)
100 estimators	0.51 (0.02)
Multi-layer perceptron	
1 × 32-dim hidden layer	0.50 (0.01)
1 × 100-dim hidden layer	0.52 (0.01)
1 × 512-dim hidden layer	0.53 (0.02)
1 × 1028-dim hidden layer	0.53 (0.02)
2 × 32-dim hidden layers	0.51 (0.02)
2 × 100-dim hidden layers	**0.55 (0.02)**

tangram shape using logit scores from CLIP. The simplest way to do this is simply taking the softmax of the logits. However, tangram shapes are outside of the training distribution for the model, perhaps explaining why it favored some images over others regardless of the content of the text.

To improve the performance of base CLIP, we trained a set of readout models to assign probabilities to images using CLIP's logits as features. Models were trained to maximize task performance (i.e., to assign high probability to the target tangram given the concatenated describer utterance). We compared four types of models: random forest, logistic regression, multi-layer perceptron (MLP), and gradient-boosted tree. Classifiers were implemented in the `scikit-learn` and `XGBoost` libraries (Chen

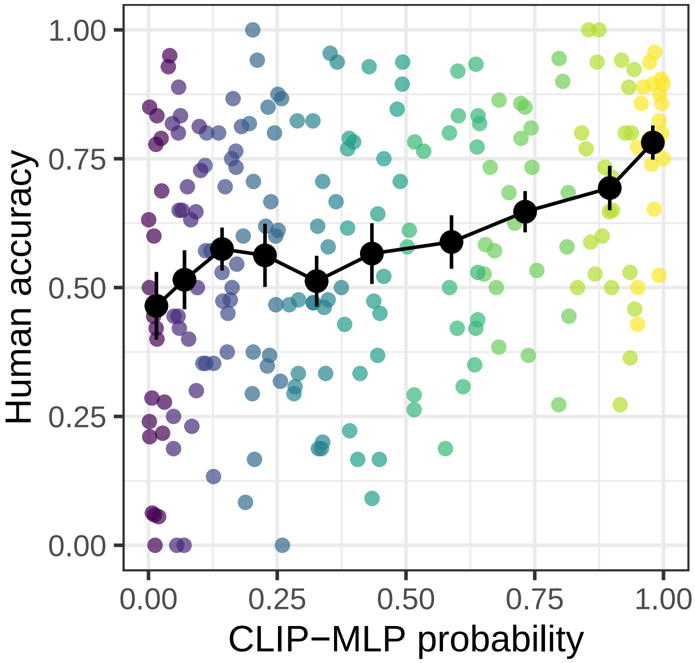


Figure 4.2: Correlation between human accuracy and CLIP-MLP probability of target in Experiment 1. Small points are individual descriptions, colored by decile of CLIP-MLP probability, large points and error bars are the bootstrapped mean and 95% CI across descriptions for each decile.

& Guestrin, 2016; Pedregosa et al., 2011). Each readout model was evaluated using 10-fold cross-validation, where the model was trained on 90% of the data and evaluated on the remaining 10%. Table 4.1 shows the cross-validated accuracy of different readout models, as well as the performance of CLIP with no readout. The MLP with two hidden layers of size 100 performed the best on held-out data; in subsequent analyses, we use the MLP trained on all the data.

4.3 Experiment 1

Our CLIP-MLP computational model was optimized for task accuracy. To validate whether this objective also results in human-like response patterns, we conducted a calibration experiment to determine if model-assigned target probabilities were aligned with the probabilities that naïve human matchers would choose particular target images across a range of utterance-target pairs.

4.3.1 Methods

We first obtained target probabilities from our CLIP-MLP model for all utterances from Boyce et al. (2024). We then used stratified sampling to select 217 trials by dividing model-predicted probabilities into deciles and choosing approximately 22 utterances per decile, spanning the 12 different possible target images. We recruited 61 participants who each saw 64 trials randomly sampled from the 217 tested trials. On average, each trial was seen by 18 participants. This experiment was pre-registered at this anonymized link.

4.3.2 Results and discussion

We obtained human accuracies on each trial by dividing the number of participants who selected the target by the total number of participants who saw the trial (Figure 4.2). There was a modest but significant positive correlation between model-predicted probabilities and human accuracies ($r = 0.33$ [0.21, 0.45]). This result suggests that model predictions were calibrated to human response patterns, albeit not perfectly. Nonetheless, the observed positive correlation suggests that our computational model carries some signal about human accuracies, validating its use in subsequent experiments as a computational comparison.

4.4 Experiment 2

As a starting point for examining the opacity of referring expressions, we focused on referring expressions from the first and last rounds of reference games. Based on the idea that conventions form across repeated communication, later-round utterances should be more opaque. To test this hypothesis, we ran a recognition experiment including descriptions from games of different sizes and communication thicknesses. Based on the patterns of cross-game similarity in Boyce et al. (2024), we expected that smaller and thicker games, whose descriptions diverged fastest, would have more idiosyncratic and opaque conventions than larger groups with thinner communication

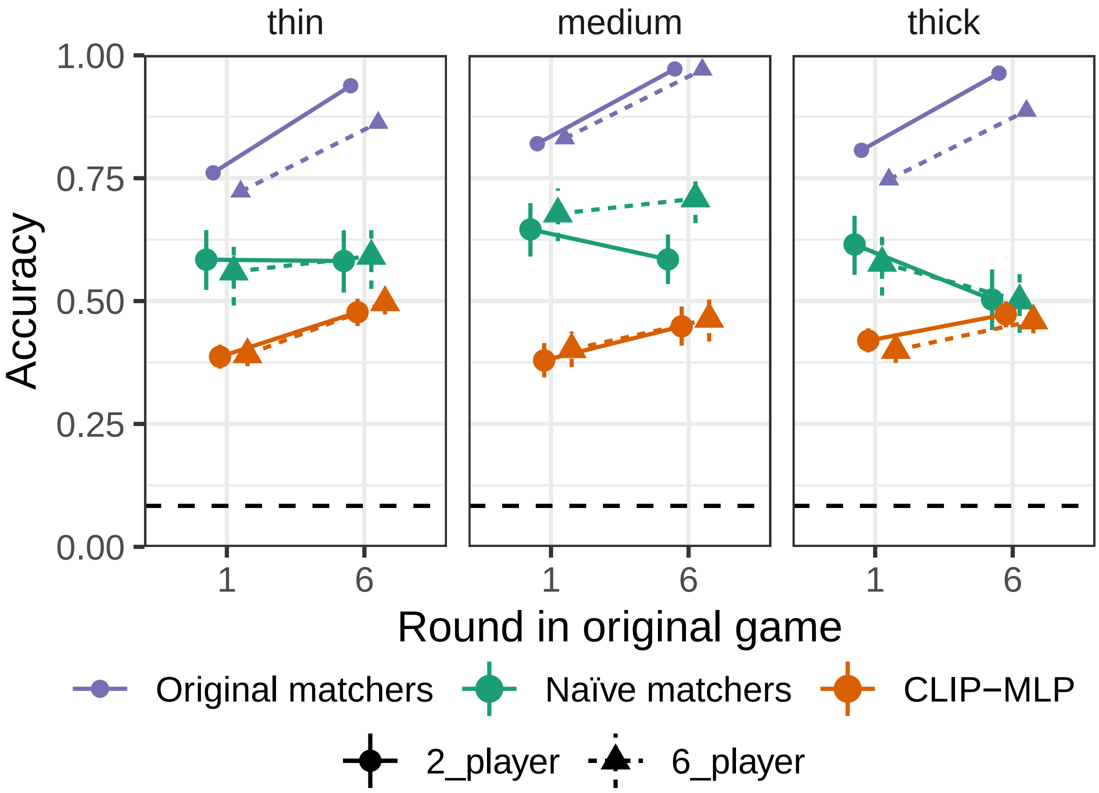


Figure 4.3: Accuracies for naïve human matchers and the CLIP-MLP model for Experiments 2a and 2b, grouped by the source of the referential description. Facets are the communication thickness of the original game and x-axis is when in the game the transcript came from. Point estimates and 95% CrI are predictions from the fixed effects of logistic and beta regressions. Bootstrapped mean accuracy from the original matchers is included as a ceiling, and random chance as a baseline.

channels.

4.4.1 Methods

Experiment 2a

To establish a baseline of how well naïve matchers could understand descriptions without context, we ran a 2 × 2 within-subjects experiment, drawing target transcripts from 2- and 6-player games from Experiment 1 of Boyce et al. (2024) and from the

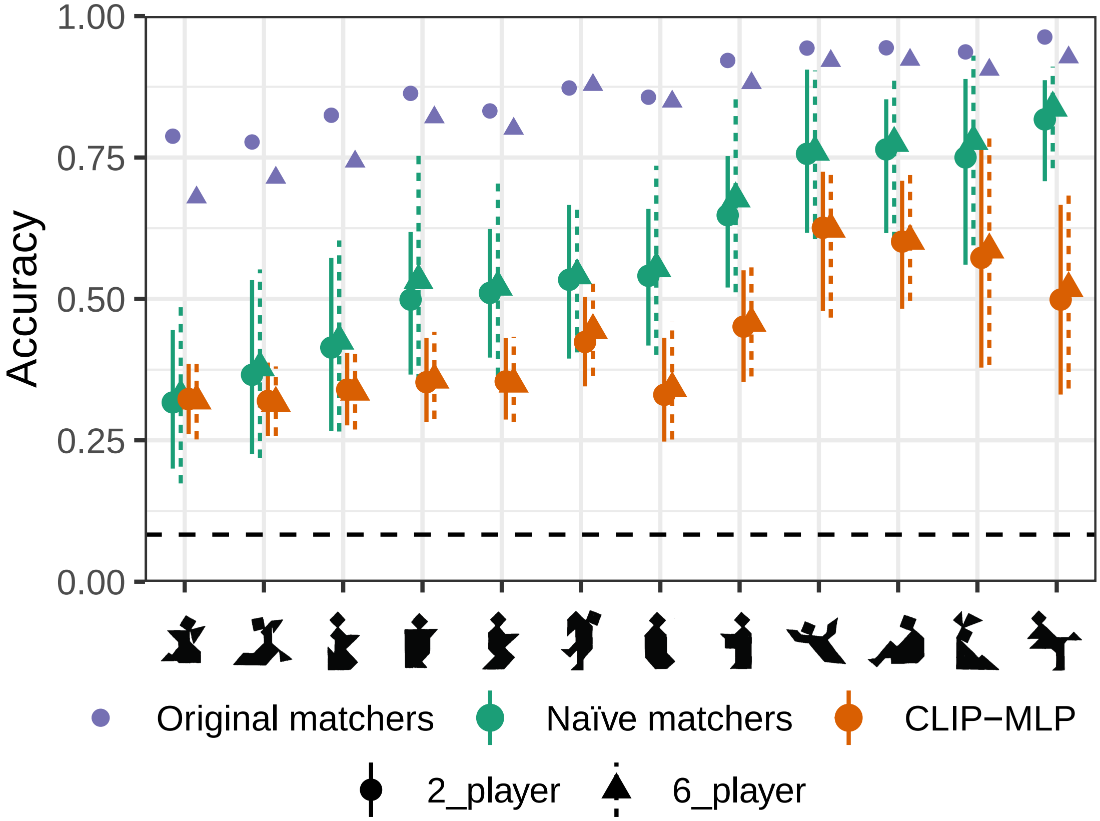


Figure 4.4: Accuracies for naïve human matchers and the CLIP-MLP model for Experiments 2a and 2b, split out by target image. Point estimates and 95% CI are predictions from the fixed effects and by-tangram random effects of logistic and beta regressions, bootstrapped across conditions. Bootstrapped mean accuracy from the original matchers is included as a ceiling, and random chance as a baseline.

first and last blocks of these games. These games had medium-thick communication channels, where matchers could send text messages to the chat, the describer role rotated each round, and matchers received limited feedback. We recruited 60 participants who each saw 60 trials (15 in each of the 4 conditions). Overall, participants saw 774 transcripts from 40 games. This experiment was pre-registered at this anonymized link.

Experiment 2b

After observing limited condition differences in Experiment 2a, we ran a follow-up experiment on descriptions from Experiment 3 of Boyce et al. (2024), where the communication channel thicknesses were more extreme. Here, we used a 2 × 2 × 2 within-subjects design, drawing our transcripts from the first and last rounds of thick and thin, 2- and 6- person games. In the "thick" condition, matchers could send text messages to the chat, one person was the describer for the whole game, and matchers received feedback on everyone's selections. In contrast, in the "thin" condition, matchers could only communicate by sending 4 emoji, the describer role rotated, and matchers recieved limited feedback. As the emoji did not have referential content, we did not include them in the transcripts shown to naïve matchers. For experiment 2b, we recruited 60 participants who each saw 64 trials (8 in each of the 8 conditions). Overall, participants saw 2392 transcripts from 163 games. This experiment was pre-registered at this anonymized link.

4.4.2 Results

Experiment 2a

For Experiment 2a, we ran a Bayesian mixed-effects logistic model of naïve matcher accuracy in `brms` (Bürkner, 2018).[1] Overall, naïve matchers were right 62% of the time, far above the 1/12 = 8.3% expected by random chance (OR = 1.93 [1.05, 3.62]). There were no large effects of condition (Figure 4.3 middle panel). Participants tended to be less accurate at descriptions from the last round (OR of last round = 0.77 [0.53, 1.10]). There was no clear effect of original group size (OR of 6-player game = 1.15 [0.89, 1.47]), but there was an interaction between round and group size (OR = 1.49 [1.06, 2.10]). Later transcripts from larger games were easier to understand, but earlier transcripts from smaller games were easier to understand. Much of the variation in accuracy was driven by the target image, which accounted for more variation than participant differences (standard deviation of image distribution = 0.98 [0.63, 1.51]; SD of participant distribution = 0.64 [0.42, 0.88]). Some images were much easier to

[1] correct ∼ group_size × round + trial_order + (group_size × round|correct_tangram) + (group_size × round + trial_order|workerid)

identify as the target than others (Figure 4.4).

Experiment 2b

For Experiment 2b, we ran a similar Bayesian mixed-effects logistic model.[2] Naïve matchers were above chance (OR = 1.81 [1.06, 3.08], Figure 4.3). As in Experiment 2a, there were not substantial effects of condition. Last-round descriptions had slightly lower accuracy (OR of last round = 0.64 [0.47, 0.85]), but there was an interaction with thickness, where for thin games, last round descriptions were less opaque (OR = 1.55 [1.02, 2.33]). Again there was strong variation based on target image (0.81 [0.51, 1.28]), which exceeded by-participant variation (0.62 [0.43, 0.83]).

Additional predictors

We considered the accuracy of the in-game matchers from Boyce et al. (2024) and the length of the description as post-hoc predictors. In both experiments, in-game accuracy was predictive of naïve matcher accuracy (Expt 2a OR = 3.33 [2.45, 4.53], Expt 2b OR = 2.39 [1.88, 3.03]). The log number of words in the description was not predictive in Experiment 2a (OR = 1.05 [0.94, 1.17]), but longer descriptions were slightly beneficial in Experiment 2b (OR = 1.10 [1.01, 1.20]).

The pattern of which conditions became more opaque in later rounds resembled the pattern of which conditions produced descriptions that diverged the most in semantic space in Boyce et al. (2024). As a post-hoc test of whether opacity might be related to semantic divergence, we used the mean semantic similarity between an utterance and other utterances in the same condition as an additional predictor of accuracy.[3] Similarity to other utterances was strongly predictive of increased accuracy in both experiments (Expt 2a: OR = 12.49 [4.70, 33.64], Expt 2b: OR = 14.75 [5.93, 36.87]) and was more predictive for the last round descriptions (Expt 2a: OR = 3.49 [1.09, 10.92], Expt 2b: OR = 4.78 [1.61, 14.25]). While exploratory, this analysis suggests that referring expressions that are further from shared semantic priors (i.e., more idiosyncratic) are harder for naïve matchers to understand.

[2] correct ∼ group_size × thickness × round + trial_order + (group_size × thickness × round|correct_tangram) + (group_size × thickness × round + trial_order|workerid)

[3] Semantic similarity was operationalized as cosine similarity between S-BERT embeddings (Reimers & Gurevych, 2019), the measure of semantic distance used in Boyce et al. (2024).

4.4.3 Model results

As a computational comparison, we used the probability the CLIP-MLP model assigned to the correct target as our dependent measure and fit a Bayesian mixed-effects beta regression on the descriptions from Experiment 2.[4] The CLIP-MLP model was far above chance, but had lower accuracy than the human participants (OR = 0.60 [0.45, 0.82]). The strongest predictor of accuracy was later round (OR = 1.32 [0.94, 1.83]), but even this was uncertain. There was substantial by-target image variation (SD = 0.46 [0.27, 0.76]).

In additional models, we checked the effect of in-game matcher accuracy, length of the description, and semantic divergence. CLIP-MLP had higher accuracy when in-game matcher accuracy was higher (OR = 1.52 [1.35, 1.71]), and when descriptions were shorter (OR for log words = 0.85 [0.82, 0.90]). The model may perform poorly on long descriptions because they are further from the model's training distribution of image captions. A description's semantic similarity to other descriptions was predictive of higher accuracy (OR = 11.85 [6.51, 21.06]), especially for last round utterances (OR = 3.46 [1.65, 7.36]), in line with the human results.

4.4.4 Discussion

Overall, naïve human matchers were fairly accurate overall, but less accurate than matchers in the original game, consistent with prior work. The computational model was less accurate, but still far above chance. The largest source of variability in accuracy was from target images, and whether earlier or later utterances were more opaque varied by game condition. The level of semantic divergence from other expressions was strongly predictive of the opacity of the expression.While this analysis was not prespecified, it still provides some suggestion that descriptions that were closer to shared semantic priors were also more interpretable.

[4] correct $\sim$ group_size $\times$ thickness $\times$ round + (group_size $\times$ thickness $\times$ round|correct_tangram)

4.5 Experiment 3

The experience of naïve matchers in Experiment 2 differed from in-game matchers in several ways; any of these could explain differences in accuracy. In-game matchers received descriptions from a consistent group, in the order they were created, and were the intended audience of the descriptions. In Experiment 3, we focused on the role of context and group-specific interaction history to tease apart some of these differences. Our primary question of interest was how much seeing the entire the conversation history in order would increase the interpretability of later round descriptions.

4.5.1 Methods

We compared naïve matchers in yoked and shuffled conditions. In the yoked condition, naïve matchers saw all the descriptions from a single game in the order they originally occurred. In the shuffled condition, naïve matchers saw all the descriptions from a single game in a randomized order.[5]

Because some descriptions are already fairly comprehensible in isolation, we focused on games that showed strong group-specificity. We hand-picked 10 games from Boyce et al. (2024) on the basis of high in-game matcher accuracy, strong patterns of descriptions shortening over repetition, and the use of idiosyncratic or non-modal referring expressions. Thus, these games showed the hallmarks of strong conventionalization to terms that were more likely to be opaque to outsiders.

We recruited 196 participants (99 in the yoked condition and 97 in shuffled) who each saw all 72 trials of one of the 10 games. This experiment was pre-registered at this anonymized link. Participants read the transcripts in a modified self-paced reading procedure where they uncovered the text word-by-word (revealed words stayed visible); only after uncovering the entire transcript could participants select an image. We do not analyze the reading time data here.

[5] For clarity, we note this is the same yoking but a different shuffling than that used in (**hawkins2023a?**).

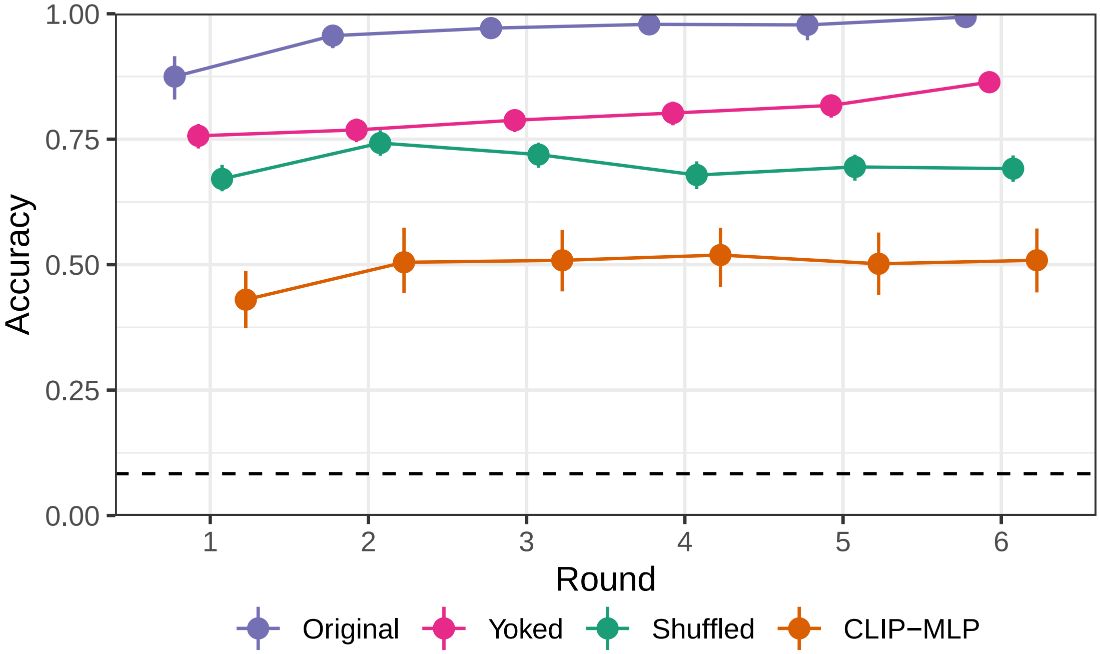


Figure 4.5: Accuracies for Experiment 3. Error bars are bootstrapped 95% CIs.

4.5.2 Results and discussion

Our primary question of interest was how much seeing the conversation history unfold in order would help participants interpret descriptions, especially those from later rounds.

We compared accuracy across the yoked and shuffled conditions with a Bayesian mixed-effects logistic regression.[6]. The descriptions were more transparent when they were presented in a yoked order (OR = 2.20 [1.63, 3.00], Figure 4.5). In the shuffled condition, there was no main effect of round number (OR for one round later = 0.99 [0.95, 1.02]), but there was a marginal interaction where the benefit of the yoked condition decreased for later rounds (OR for one round later = 0.94 [0.89, 1.00]). This was offset by matchers in both conditions improving at the task over time (OR for one trial later in matcher viewing order = 1.02 [1.02, 1.02]).

Comparing to the performance of in-game matchers, we separated out the benefits

[6] correct ∼ orig_repNum × condition + matcher_trialNum + (1|gameId) + (1|correct_tangram) + (1|workerid)

of seeing the descriptions in order versus being a participant in the group.[7] There was a benefit to seeing the items in order (OR = 2.24 [1.63, 3.04]) and a larger benefit to being a participant during the game (OR = 4.35 [2.77, 6.89]). The benefit of seeing the items in order waned in later blocks (OR = 0.94 [0.89, 1.00]), but the benefit of being in the game did not (OR = 1.06 [0.95, 1.18]). In all cases, there was a baseline improvement over trials (OR = 1.02 [1.02, 1.02]). As a caveat, we note that in-game matchers and naïve matchers may have varied from each other in terms of effort and time spent on the task, and thus the comparison should be interpreted cautiously.

The accuracy of the CLIP-MLP model was worse than the shuffled human results, and did not change across rounds (OR for one round later = 1.02 [0.97, 1.07]). The larger difference between naïve human and CLIP-MLP accuracies in Experiment 3 than Experiment 2 suggests that the shuffled ordering still provides useful context that helps matchers understand the conventions. This history was not available to the CLIP-MLP model which saw every description as a one-shot task.

4.6 General Discussion

Real-world conventions vary in whether they are opaque to outsiders (“rizz”) or interpretable even to those who don’t produce them (“roundabout”). Convention formation in the real world is difficult to study, so iterated reference games are a method for operationalizing convention formation for experimental study. In reference games, conventions are partner-specific: different groups’ evolving conventions follow different paths through semantic space. Despite the number of studies on iterated reference games, few studies have examined whether the descriptions are interpretable by outsiders.

Across multiple experiments, we found that naïve human matchers were far above chance at identifying the targets, and our computational model was also above chance. For both humans and models, more variation was explained by the target image than the round or game condition the descriptions came from, suggesting that conventionalization was not the primary driver of how difficult an expression was to interpret.

[7] correct $\sim$ orig_repNum $\times$ order + orig_repNum $\times$ setting + matcher_trialNum + (1|gameId) + (1|correct_tangram) + (1|workerid)

Even for games selected for strong conventionalization, naïve matchers had high accuracy overall, although this accuracy was further increased if they saw the conversation history in order. Exploratory analyses also suggested that more idiosyncratic descriptions were more opaque. Our findings are consistent with a lexical uncertainty approach, where expressions that are closer to overall priors are easier to understand, and groups that have fewer people and thicker channels are more able to break away from these priors and have conventions that drift farther apart in semantic space.

Limiting the generality of our findings, our experimental and computational results were only on a specific set of iterated reference game transcripts and images. Some images may lend themselves to more transparent descriptions because they are more iconic, with a narrower prior over different ways they could be conceptualized, or they may be further from competitors within this pool of images. Future work sampling across larger sets of images (such as Ji et al., 2022) could probe image-level factors.

Nonetheless, this work has demonstrated the utility of adopting a broader perspective on convention comprehension. In particular, the use of computational modelling allowed for a means to estimate semantics under lexical uncertainty without requiring symbolic semantic representations. Future work could capitalize on this approach to better understand semantic dynamics underlying convention formation, as well as provide further quantitative investigations of pragmatics models like RSA and CHAI. These directions will help us to better understand the nature of reference and conventions, and how humans navigate the complex and ever-evolving landscape of communication.

Conclusion

Wow at some point there will be text that goes here. Isn't that exciting!

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

Ahern, T. C. (1994). The effect of interface on the structure of interaction in computer-mediated small-group discussion. *Journal of Educational Computing Research*, *11*(3), 235–250.

Almaatouq, A., Becker, J., Houghton, J. P., Paton, N., Watts, D. J., & Whiting, M. E. (2020). Empirica: A virtual lab for high-throughput macro-level experiments. *arXiv:2006.11398 [Cs]*. https://arxiv.org/abs/2006.11398

Almaatouq, A., Griffiths, T. L., Suchow, J. W., Whiting, M. E., Evans, J., & Watts, D. J. (2024). Beyond playing 20 questions with nature: Integrative experiment design in the social and behavioral sciences. *Behavioral and Brain Sciences*, *47*, e33. https://doi.org/10.1017/S0140525X22002874

Bates, E. (1974). Acquisition of pragmatic competence. *Journal of Child Language*, *1*(2), 277–281. https://doi.org/10.1017/S0305000900000702

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

Bohn, M., & Frank, M. C. (2019). *The pervasive role of pragmatics in early language.* https://doi.org/10.31234/osf.io/v8e56

Bohn, M., Kachel, G., & Tomasello, M. (2019). Young children spontaneously recreate core properties of language in a new modality. *Proceedings of the National Academy of Sciences*, *116*(51), 26072–26077. https://doi.org/10.1073/pnas.1904871116

Bohn, M., Schmidt, L. S., Schulze, C., Frank, M. C., & Tessler, M. H. (2022). Modeling Individual Differences in Children's Information Integration During Pragmatic Word Learning. *Open Mind*, *6*, 311–326. https://doi.org/10.1162/opmi_a_00069

Boyce, V., Hawkins, R. D., Goodman, N. D., & Frank, M. C. (2024). Interaction structure constrains the emergence of conventions in group communication. *Proceedings of the National Academy of Sciences*, *121*(28), e2403888121. https://doi.org/10.1073/pnas.2403888121

Branigan, H. (2006). Perspectives on multi-party dialogue. *Research on Language and Computation*, *4*(2), 153–177.

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

Bürkner, P.-C. (2018). Advanced bayesian multilevel modeling with the r package brms. *The R Journal*, *10*(1), 395–411.

Caplow, T. (1957). Organizational size. *Administrative Science Quarterly*, 484–505.

Carletta, J., Garrod, S., & Fraser-Krauss, H. (1998). Placement of Authority and Communication Patterns in Workplace Groups: The Consequences for Innovation. *Small Group Research*, *29*(5), 531–559. https://doi.org/10.1177/1046496498295001

Carruthers, P. (2013). Mindreading in Infancy. *Mind & Language*, *28*(2), 141–172. https://doi.org/10.1111/mila.12014

Cazden, C. B. (1988). *Classroom discourse: The language of teaching and learning.* ERIC.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–794.

Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, *50*(1), 62–81.

Clark, H. H., & Schaefer, E. F. (1987). Concealing one's meaning from overhearers. *Journal of Memory and Language*, *26*(2), 209–225. https://doi.org/10.1016/0749-596X(87)90124-0

Clark, H. H., & Wilkes-Gibbs, D. (1986). *Referring as a collaborative process.*

Cohn-Gordon, R., Goodman, N., & Potts, C. (2018). Pragmatically Informative Image Captioning with Character-Level Inference. *arXiv:1804.05417 [Cs]*. https://arxiv.org/abs/1804.05417

Cohn-Gordon, R., Levy, R., & Bergen, L. (2019). *The pragmatics of multiparty communication.*

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*.

Dewhirst, H. D. (1971). Influence of perceived information-sharing norms on communication channel utilization. *Academy of Management Journal*, *14*(3), 305–315.

Fay, N., Garrod, S., & Carletta, J. (2000). Group Discussion as Interactive Dialogue or as Serial Monologue: The Influence of Group Size. *Psychological Science*, *11*(6), 481–486. https://doi.org/10.1111/1467-9280.00292

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. *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). *Alcohol and Common Ground: The Effects of Intoxication on Linguistic Markers of Shared Understanding during Social Exchange*. OSF Preprints. `https://doi.org/10.31219/osf.io/xrw6z`

Garrod, S., Fay, N., Lee, J., Oberlander, J., & MacLeod, T. (2007). Foundations of Representation: Where Might Graphical Symbol Systems Come From? *Cognitive Science*, *31*(6), 961–987. `https://doi.org/10.1080/03640210701703659`

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`

Ginzburg, J., & Fernandez, R. (2005). *Action at a distance: The difference between dialogue and multilogue*. 9.

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. `https://www.jstor.org/stable/23082551`

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

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`

Graham, S. A., Sedivy, J., & Khu, M. (2014). That's not what you said earlier:

Preschoolers expect partners to be referentially consistent. *Journal of Child Language*, *41*(1), 34–50. https://doi.org/10.1017/S0305000912000530

Grand, G., Blank, I. A., Pereira, F., & Fedorenko, E. (2022). Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature Human Behaviour*, *6*(7), 975–987. https://doi.org/10.1038/s41562-022-01316-8

Grigoroglou, M., & Papafragou, A. (2019). Interactive contexts increase informativeness in children's referential communication. *Developmental Psychology*, *55*(5), 951–966. https://doi.org/10.1037/dev0000693

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

Gul, M. O., & Artzi, Y. (2024). *CoGen: Learning from Feedback with Coupled Comprehension and Generation* (No. arXiv:2408.15992). arXiv. https://doi.org/10.48550/arXiv.2408.15992

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

Haber, J., Baumgärtner, T., Takmaz, E., Gelderloos, L., Bruni, E., & Fernández, R. (2019). The PhotoBook Dataset: Building Common Ground through Visually-Grounded Dialogue. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1895–1910. https://doi.org/10.18653/v1/P19-1184

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., Franke, M., Frank, M. C., Goldberg, A. E., Smith, K., Griffiths, T. L., & Goodman, N. D. (2021). *From partners to populations: A hierarchical Bayesian account of coordination and convention* (No. arXiv:2104.05857). arXiv. https://doi.org/10.48550/arXiv.2104.05857

Hawkins, R. D., Franke, M., Frank, M. C., Goldberg, A. E., Smith, K., Griffiths, T. L., & Goodman, N. D. (2023). From partners to populations: A hierarchical Bayesian account of coordination and convention. *Psychological Review*, *130*(4), 977.

Hawkins, R. D., Kwon, M., Sadigh, D., & Goodman, N. D. (2020). Continual adaptation for efficient machine communication. *arXiv:1911.09896 [Cs]*. `https://arxiv.org/abs/1911.09896`

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*.

Hiltz, S. R., Johnson, K., & Turoff, M. (1986). Experiments in group decision making: Communication process and outcome in face-to-face versus computerized conferences. *Human Communication Research*, *13*(2), 225–252.

Horton, W. S., & Gerrig, R. J. (2002a). SpeakersÕ experiences and audience design: Knowing when and knowing how to adjust utterances to addresseesq. *Journal of Memory and Language*, 18.

Horton, W. S., & Gerrig, R. J. (2002b). 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`

Huang, Y. T., & Snedeker, J. (2009). Online interpretation of scalar quantifiers: Insight into the semantics–pragmatics interface. *Cognitive Psychology*, *58*(3), 376–415. `https://doi.org/10.1016/j.cogpsych.2008.09.001`

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`

Ji, A., Kojima, N., Rush, N., Suhr, A., Vong, W. K., Hawkins, R., & Artzi, Y. (2022). Abstract Visual Reasoning with Tangram Shapes. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 582–601). Association for Computational Linguistics. `https://doi.org/10.18653/v1/2022.emnlp-main.38`

Kang, Y., Wang, T., & de Melo, G. (2020). Incorporating Pragmatic Reasoning Communication into Emergent Language. *Advances in Neural Information Processing Systems*, *33*, 10348–10359.

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

Keen, R. (2003). Representation of Objects and Events: Why Do Infants Look So Smart and Toddlers Look So Dumb? *Current Directions in Psychological Science*, *12*(3), 79–83. `https://doi.org/10.1111/1467-8721.01234`

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., & Bricker, P. D. (1967). Effects of transmission delay and access delay on the efficiency of verbal communication. *The Journal of the Acoustical Society of America*, *41*(2), 286–292.

Krauss, R. M., Garlock, C. M., Bricker, P. D., & McMahon, L. E. (1977). The role of audible and visible back-channel responses in interpersonal communication. *Journal of Personality and Social Psychology*, *35*(7), 523.

Krauss, R. M., & Glucksberg, S. (1969). The Development of Communication: Competence as a Function of Age. *Child Development*, *40*(1), 255–266. `https:`

//doi.org/10.2307/1127172

Krauss, R. M., & Weinheimer, S. (1964). Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*, *1*(1-12), 113–114. https://doi.org/10.3758/BF03342817

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

Kraut, R. E., Lewis, S. H., & Swezey, L. W. (1982). Listener responsiveness and the coordination of conversation. *Journal of Personality and Social Psychology*, *43*(4), 718.

Le, H., Daryanto, T., Zhafransyah, F., Wijaya, D., Coppock, E., & Chin, S. (2022). *Referring Expressions with Rational Speech Act Framework: A Probabilistic Approach* (No. arXiv:2205.07795). arXiv. https://doi.org/10.48550/arXiv.2205.07795

Leeuw, J. R. de, Gilbert, R. A., & Luchterhandt, B. (2023). jsPsych: Enabling an Open-Source Collaborative Ecosystem of Behavioral Experiments. *Journal of Open Source Software*, *8*(85), 5351. https://doi.org/10.21105/joss.05351

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

Leung, A., Yurovsky, D., & Hawkins, R. D. (2024). Parents spontaneously scaffold the formation of conversational pacts with their children. *Child Development*, *n/a*(n/a). https://doi.org/10.1111/cdev.14186

Lewis, D. (1969). *Convention: A philosophical study*. John Wiley & Sons.

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

MacMillan, J., Entin, E. E., & Serfaty, D. (2004). Communication overhead: The hidden cost of team cognition. In *Team cognition: Understanding the factors that drive process and performance.* (pp. 61–82). American Psychological Association. https://doi.org/10.1037/10690-004

Matthews, D., Butcher, J., Lieven, E., & Tomasello, M. (2012). Two- and Four-Year-Olds Learn to Adapt Referring Expressions to Context: Effects of Distracters and Feedback on Referential Communication. *Topics in Cognitive Science*, *4*(2), 184–210. `https://doi.org/10.1111/j.1756-8765.2012.01181.x`

Matthews, D., Lieven, E., Theakston, A., & Tomasello, M. (2006). The effect of perceptual availability and prior discourse on young children's use of referring expressions. *Applied Psycholinguistics*, *27*(3), 403–422. `https://doi.org/10.1017/S0142716406060334`

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`

Misyak, J., Noguchi, T., & Chater, N. (2016). Instantaneous Conventions: The Emergence of Flexible Communicative Signals. *Psychological Science*, *27*(12), 1550–1561. `https://doi.org/10.1177/0956797616661199`

Monroe, W., & Potts, C. (2015). Learning in the Rational Speech Acts Model. *arXiv:1510.06807 [Cs]*. `https://arxiv.org/abs/1510.06807`

Morisseau, T., Davies, C., & Matthews, D. (2013). How do 3- and 5-year-olds respond to under- and over-informative utterances? *Journal of Pragmatics*, *59*, 26–39. `https://doi.org/10.1016/j.pragma.2013.03.007`

Murfitt, T., & McAllister, J. (2001). The Effect of Production Variables in Monolog and Dialog on Comprehension by Novel Listeners. *Language and Speech*, *44*(3), 325–350. `https://doi.org/10.1177/00238309010440030201`

Muttenthaler, L., & Hebart, M. N. (2021). THINGSvision: A Python Toolbox for Streamlining the Extraction of Activations From Deep Neural Networks. *Frontiers in Neuroinformatics*, *15*. `https://doi.org/10.3389/fninf.2021.679838`

Nadig, A. S., & Sedivy, J. C. (2002). Evidence of Perspective-Taking Constraints in Children's On-Line Reference Resolution. *Psychological Science*, *13*(4), 329–336. https://doi.org/10.1111/j.0956-7976.2002.00460.x

Nilsen, E. S., & Graham, S. A. (2009). The relations between children's communicative perspective-taking and executive functioning. *Cognitive Psychology*, *58*(2), 220–249. https://doi.org/10.1016/j.cogpsych.2008.07.002

Noveck, I. A. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition*, *78*(2), 165–188. https://doi.org/10.1016/S0010-0277(00)00114-1

Ohmer, X., Franke, M., & König, P. (2022). Mutual Exclusivity in Pragmatic Agents. *Cognitive Science*, *46*(1), e13069. https://doi.org/10.1111/cogs.13069

Parisi, J. A., & Brungart, D. S. (2005). Evaluating communication effectiveness in team collaboration. *Ninth European Conference on Speech Communication and Technology (INTERSPEECH)*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

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

Potts, C., Lassiter, D., Levy, R., & Frank, M. C. (2015). Embedded Implicatures as Pragmatic Inferences under Compositional Lexical Uncertainty. *Journal of Semantics*, ffv012. https://doi.org/10.1093/jos/ffv012

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

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). *Learning Transferable Visual Models From Natural Language Supervision* (No. arXiv:2103.00020). arXiv. `https://doi.org/10.48550/arXiv.2103.00020`

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust Speech Recognition via Large-Scale Weak Supervision.*

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`

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks* (No. arXiv:1908.10084). arXiv. `https://doi.org/10.48550/arXiv.1908.10084`

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`

San Juan, V., Khu, M., & Graham, S. A. (2015). A New Perspective on Children's Communicative Perspective Taking: When and How Do Children Use Perspective Inferences to Inform Their Comprehension of Spoken Language? *Child Development Perspectives*, *9*(4), 245–249. `https://doi.org/10.1111/cdep.12141`

Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology*, *21*(2), 211–232. `https://doi.org/10.1016/0010-0285(89)90008-X`

Schuster, S., & Degen, J. (2020). I know what you're probably going to say: Listener adaptation to variable use of uncertainty expressions. *Cognition*, *203*, 104285. `https://doi.org/10.1016/j.cognition.2020.104285`

Seaman, C. B., & Basili, V. R. (1997). Communication and organization in software development: An empirical study. *IBM Systems Journal*, *36*(4), 550–563.

Swaab, R. I., Galinsky, A. D., Medvec, V., & Diermeier, D. A. (2012). The communication orientation model: Explaining the diverse effects of sight, sound, and synchronicity on negotiation and group decision-making outcomes. *Personality and Social Psychology Review*, *16*(1), 25–53.

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`

Tannen, D. (2005). *Conversational style: Analyzing talk among friends*. Oxford University Press.

Tolins, J., & Fox Tree, J. E. (2016). Overhearers Use Addressee Backchannels in Dialog Comprehension. *Cognitive Science*, *40*(6), 1412–1434. `https://doi.org/10.1111/cogs.12278`

Tomasello, M. (2008). *Origins of human communication*. MIT press.

Traum, D. (2004). Issues in Multiparty Dialogues. In G. Goos, J. Hartmanis, J. van Leeuwen, & F. Dignum (Eds.), *Advances in Agent Communication* (Vol. 2922, pp. 201–211). Springer Berlin Heidelberg. `https://doi.org/10.1007/978-3-540-24608-4_12`

Turan-Küçük, E. N., & Kibbe, M. M. (2024). Three-year-olds' ability to plan for mutually exclusive future possibilities is limited primarily by their representations of possible plans, not possible events. *Cognition*, *244*, 105712. `https://doi.org/10.1016/j.cognition.2023.105712`

Weber, R. A., & Camerer, C. F. (2003). Cultural Conflict and Merger Failure: An Experimental Approach. *Management Science*, *49*(4), 16.

White, J., Mu, J., & Goodman, N. D. (2020). Learning to refer informatively by amortizing pragmatic reasoning. *arXiv:2006.00418 [Cs]*. `https://arxiv.org/abs/2006.00418`

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

Wittgenstein, L. (1953). *Philosophical investigations*. Wiley-Blackwell.

Yoon, E. J., Frank, M. C., Tessler, M. H., & Goodman, N. D. (2018). *Polite speech emerges from competing social goals* [Preprint]. PsyArXiv. `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

Zack, M. H. (1993). Interactivity and communication mode choice in ongoing management groups. *Information Systems Research*, *4*(3), 207–239.

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.*