Communicative reduction in referring expressions within a multi-player negotiation game

Veronica Boyce (vboyce@stanford.edu)
Department of Psychology, Stanford University

Michael C. Frank (mcfrank@stanford.edu)
Department of Psychology, Stanford University

Abstract

The ability to form novel conventions is a key signature of efficient linguistic communication. Reduction of referring expressions, one measure of convention formation, is found robustly in dyadic repeated reference games when the target images are initially difficult to name. In reference games, participants share the explicit goal of establishing joint reference. However, establishing reference is a key subgoal of many conversations where interlocuters have more complex goals. In the current work, we analyze a dataset where reference was embedded in strategic 3- player negotiation and coordination games. In these more complex games, we found that the patterns of reduction and convergence to conventions still held across two different incentive conditions, with some modest differences between the conditions.

Keywords: Convention formation; reference games; reduction of referring expressions; multi-party communication

Successful communication is grounded in a shared understanding of what utterances mean in the context of their production. Whether the communicative goals are cooperative or adversarial, interlocutors need to establish joint reference to be effective. In many cases, conventional word meanings are enough, but there are also contexts where objects to be referred to are not easy to distinguish. In these situations, establishing reference is more difficult, but no less important.

The formation of ad-hoc referring expressions has been studied extensively in the context of dyadic reference games (Clark & Wilkes-Gibbs, 1986; Krauss & Weinheimer, 1964). In these, two participants see a set of images (often abstract shapes) and one person is tasked with identifying an image so the other person can pick it out. People can be very successful at this task, achieving high levels of matches. Over repeated reference to the same images, pairs develop shorthand names for the images, leading to shorter descriptions in later repetitions. These nicknames conventionalize within each pair, as pairs tend to stick to the same description, but nicknames are idiosyncratic and differ between pairs. This task has been a well-studied microcosm for understanding reference which has proven useful for testing theories about how referring expressions originate, how expressions are designed, and when they change (Brennan & Clark, 1996; Leung, Hawkins, & Yurovsky, 2020; Metzing & Brennan, 2003; Weber & Camerer, 2003; Yoon & Brown-Schmidt, 2019).

The implicit theoretical claim is that reduction to partner-specific conventions occurs *whenever* people interact repeatedly in ways that require reference to some objects without adequate conventional names. Indeed, reduction does occur regardless of modality, having been found in both oral and written communication (Brennan & Clark, 1996; Hawkins, Frank, & Goodman, 2020). But do these phenomena generalize beyond artificial reference game environments?

Our current work begins to address this question by looking for reduction phenomena in a slightly more naturalistic setting. The conversational goals of reference games are not representative of typical language use. In everyday communication, establishing reference is often an instrumental subgoal in a larger conversation. The overall task may be cooperative, such as asking for kitchen equipment while cooking together, or adversarial, such as identifying items while negotiating a division of assets. In contrast, in the reference game context, the goal is explicitly cooperative: match corresponding images.

We build on work by Mankewitz et al. (2021), who created a task with the aim of bridging the gap between reference games with hard-to-name objects and negotiation datasets with easily nameable objects. They created a 3-person game where players each selected what flowers to grow from a set of 6 flowers with variable and partially hidden value. Players got the value of the flower only if they were the only one to select it, which incentivized coordination and negotiation. However, all the flowers shown were the same color, so they were not easily nameable, and the flowers repeated, leading to opportunities for ad-hoc conventionalization. This game structure encouraged players to describe flowers while never explicitly telling a player to refer to any specific flower. The experiment had two conditions, one where players within a group won points together (and thus had fully aligned incentives) and another where they won points as individuals (and didn't). Mankewitz et al. (2021) found a slight decrease in language over the course of the games, but they did not conduct a detailed analysis to assess whether the decreases in words were associated with shorter and more conventionalized referring expressions. However, this pattern of decreased language production is actually totally consistent with increased task familiarity, fatigue, and/or less communication about overall strategy, with no change in the actual referring language. Here, we address this gap by separating out the referring expressions in order to check for patterns of convention-formation.

Our work addresses whether participants in a complex,

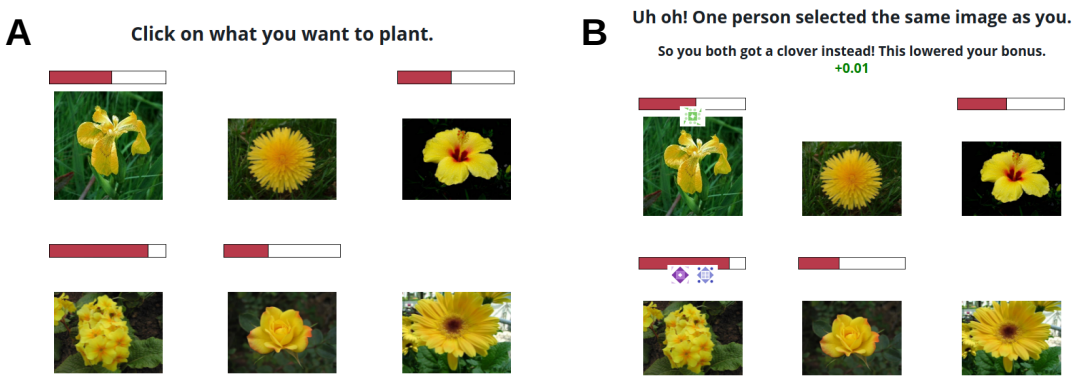

Figure 1: Player interface. Panel A shows the selection phase, where each participant sees 6 flowers, 4 with value bars. Panel B shows the feedback stage, after all players have selected. Players see player icons on the value bars indicating who chose what. When multiple players select the same flower, they recieve a lower value rather than what is shown.

multi-player strategic game with differing incentives and shared knowledge are able to develop conventionalized reference expressions. We explore whether the phenomena found in classic dyadic iterated reference games can also be found in this more naturalistic domain, and whether the results differ across the individual and shared incentive conditions. The key questions we addressed were:

1. Did referring expressions reduce over time?
2. How did referring expressions change within and between groups over the course of the games?
3. How did different incentive conditions influence referring expression formation?

Historically, studies of convention formation have had to rely on proxies for convention formation such as reduction, subjective judgments of similarity, and manual measures of lexical overlap. In contrast, we use natural language processing tools for mapping sentences to semantic vector spaces to quantify semantic similarities between sentences, following recent work by Hawkins et al. (2020) and Boyce, Hawkins, Goodman, & Frank (2022).

Methods

This paper presents a reanalysis of data from Mankewitz et al. (2021). We describe both the original data collection and the additional data processing done for reanalysis. The materials, data, and code are all available at `https://github.com/vboyce/AA-flowers`.

Procedure

As described in Mankewitz et al. (2021), participants played a real-time coordination game in groups of three, implemented using Empirica (Almaatouq et al., 2020). On each trial, each group saw a set of 6 flower images (Figure 1A) . Each participant saw the values for 4 of the flowers (represented as a colored bar), such that each flower's value was hidden from one participant. Players could coordinate and discuss using a chat box before each player selected a flower. There were no restrictions on what participants could say in the chat.

If one player selected a flower, it was worth the shown reward; if multiple players collided and selected the same flower, they each got a lower reward instead (Figure 1B). Players needed to communicate about the flowers in order to coordinate their selections. The rewards translated into a monetary bonus for the participants at the end of the game.

In *individual utility* games, each player earned points for the flowers they selected; in the *shared utility* games, the points were averaged together, and all players in a game got the same reward. This made for slightly different incentives; in an individual game, players wanted to maximize the rewards of flowers they selected, and only cared about avoiding collisions with other players' selections; in a shared game, players wanted their teammates to select different high reward flowers, and were indifferent on who selected the highest one.

Each game was assigned a color of flower (white, red, yellow, purple) and the flower images were drawn from a set of 12 for that color, so players saw the same flowers repeatedly across the game, in different combinations. Each game consisted of 24 trials. Different flowers of the same color created situations where players did not have established names for the flowers in context and needed to develop shared referring expressions to clearly communicate with their partners.

After the game, players were asked how they would describe each of the images to their teammates.

Participants

Mankewitz et al. (2021) recruited 150 participants total for 25 3-person games in each of the individual and shared utilities conditions. Games took roughly 20 minutes with wide variability (5-40 minutes). If a participant disconnected, the game ended, so some games did not complete the full 24 trials as participants dropped out early. We excluded incomplete games from analyses, which left us with 18 games in

the individual utilities condition and 21 in the shared utilities condition (117 participants included in the analyses).

Textual annotations

We annotated the chat transcripts to extract all referring expressions and identify which flower image each expression referred to. We corrected the referring expressions for spelling errors. Annotations were done primarily by the first author, with some done by a research assistant.

Two games consisting of 121 utterances were annotated by both annotators. 117 utterances were identified as containing reference expressions by at least one annotator; of these, the annotators agreed on the exact reference expression for 105 (90%) of the cases. The second annotator only coded the target of the referring expression for 60 of the utterances. Of these, the two annotators agreed on the target in 59 (98%) of cases. We take this level of inter-annotator agreement as an indication that the reference spans and targets were identified reliably. We extracted a total of 3395 referring expressions.

SBERT embeddings

Following recent work by Hawkins et al. (2020) and Boyce et al. (2022), we used tools for natural language processing to quantify similarity between phrases in order to assess the dynamics of convergence to conventions.

We embedded each of the extracted referring expressions using SBERT (Reimers & Gurevych, 2019) which maps each utterance to a vector in a high-dimensional semantic space. We used cosine distance between pairs of vector embeddings as a measure of similarity between the corresponding referring expressions. Higher cosine values correspond to vectors close together, and thus, reference expressions with more similar meanings. Some examples of pairs of referring expressions and their similarities are shown in Table 2.

For pairwise comparisons between referring expressions produced during the game, we smoothed out the random sampling and sparseness of which flowers were presented on the same trial by comparing descriptions with those from the same trial and the two later trials (thus the similarity at trial 10 comes from comparisons of trial 10 to trials 10, 11, and 12).

Analytic methods

All models were run using rstan (Stan Development Team, 2023). We coded condition as 0 for individual utilities (baseline condition) and 1 for shared utilities.

Results

We address each of the questions of interest in order: 1) did referring expressions reduce over time, 2) how referring expressions changed during games, and 3) how the incentive manipulation influenced referring expressions.

Reduction of referring expressions

In each trial of the game, there were 6 flowers visible on the screen, and participants wanted to each pick a different one. In general, each trial contained references to 2 or 3 distinct flowers: individual utilities: mean of 2.3 (sd: 1.34), shared utilities: 2.33 (sd: 1.12). Most players referred to one flower each trial: individual utilities: 1.04 (sd: 0.92), shared utilities: 0.97 (sd: 0.78).

Participants employed a range of reference strategies, including referring to the flowers by a common name (ex. rose, iris), referring to the number or groupings of flowers, referring to properties of the flowers (ex. spiky), referring to background elements (ex. white house), and analogies to other objects (ex. flame). Descriptions sometimes combined multiple of these approaches.

The amount of referring language decreased over the course of the game, consistent with the dyadic reference game pattern. As shown in Figure 2A, the number of words of referring language decreased across the game (trial: -0.154 CrI=[-0.165, -0.144]). The reduction in referring language was driven by shortening referring expressions (Figure 2B), while the total number of referring expressions per trial remained constant (Figure 2C). See Table 1, flowers 1 and 2, for examples of reduction.

Convergence of referring expressions

The classic reference game phenomena of convergence to partner-specific referring expressions can be quantified using similarity metrics as the combination of three patterns occurring over time: 1) within a group, descriptions of the same flower converge, 2) within a group, descriptions of different flowers diverge, and 3) between groups, descriptions of the same flower diverge. We assess all of these in two ways: 1) comparing referring expressions within the game to other referring expressions from the same time to get at the dynamic changes, and 2) comparing referring expressions within the game to the names players gave the flowers after the game, to track how expressions evolve towards these final conventions.

Within games Within games, the patterns from dyadic reference games predict that similarity for descriptions of the same image will increase over time while similarity for descriptions of different images will decrease over time. Table 1 offers examples of how some descriptions changed over time, and Table 2 gives similarity measures between selected pairs of these descriptions.

We found that for descriptions of the same flower, cosine similarity increased over time (trial: 0.009 CrI=[0.007, 0.011], see Figure 3 upper left and middle panels). Utterances were more similar if they were produced by the same participant (0.062 CrI=[0.02, 0.104]), although this did not interact with trial number (-0.002 CrI=[-0.006, 0.001]).

For descriptions of different flowers, cosine similarity decreased over time (trial: -0.006 CrI=[-0.007, -0.006], see Figure 3 lower left and middle panels). Utterances being produced by the same person did not have an effect (0.003 CrI=[-0.008, 0.014]) or an interaction with trial number (0 CrI=[-0.001, 0.001]).

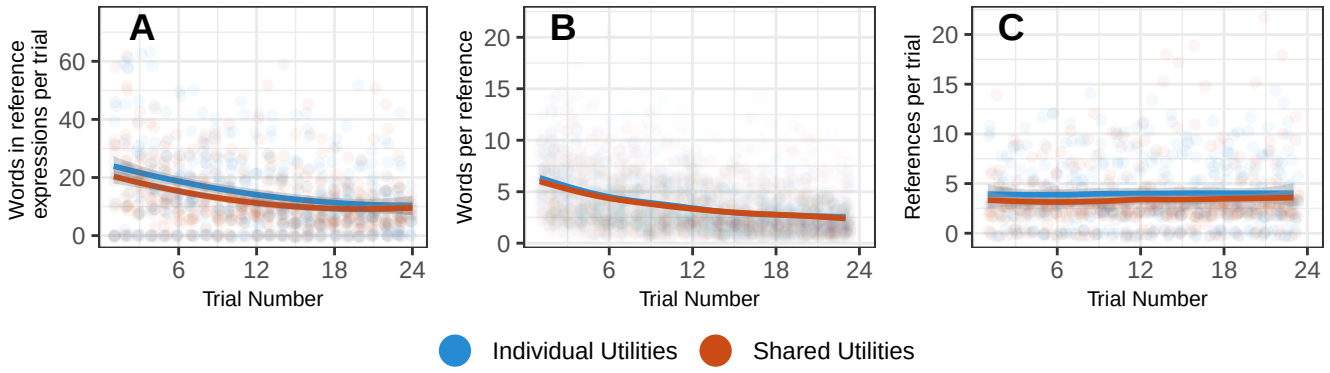

Figure 2: Words produced in referring expressions across trials in both conditions. A: Total words of referring language per game per trial. B: Words in each reference expression by trial. C: Total number of reference expressions per game per trial.

Table 1: Examples of descriptions of different flowers from different games, illustrating reduction and convergence phenomena. Games are numbered, letters refer to players within games. Images of flowers can be found in Figure 1: flower 1 upper left, flower 2 lower center, flower 3 lower left.

Flower	Game	Trial	Expression
1	1A	2	not sure what kind of flower it is but the droopy-ish one
1	1C	4	droopy iris flower
1	1B	21	droopy
2	2B	2	the red middle with spike
2	2B	3	the red center
2	2A	20	red middle
2	3C	6	the one with the dark red centre
2	3A	13	the one with black background
2	3A	24	black background
3	1A	4	the big cluster of flowers with the orange in the middle
3	1A	23	cluster
3	2C	24	bundle
3	3B	24	multi flowers

These results replicate and generalize the pattern found in prior literature: convergence in expressions for the same referent games and divergence between referents (Hawkins et al., 2020).

Between games The theory of partner-specificity predicts that descriptions to the same flower in different groups will diverge over time as each group forms their own distinct convention. Table 1 shows how descriptions of the same flower may diverge (flower 2), but also how groups might choose related but different descriptions (flower 3).

For descriptions of the same flower across games, we found that cosine similarities decreased over time (trial: -0.005 CrI=[-0.006, -0.004], see Figure 3 upper right panel). This between-game decrease is also consistent with prior reports in more restricted reference games.

Comparison with post-game convention reports After the game, players provided the description they would use to identify each flower to their teammates. We treated these descriptions as the conventionalized names and looked at how the in-game descriptions developed into these conventions.

Referring descriptions later in a game were more similar to the convention than descriptions earlier in the game (trial: 0.006 CrI=[0.005, 0.007], see Figure 5 upper panels). Utterances were more similar to the convention given by the same person (versus the convention given by a groupmate) (0.053 CrI=[0.023, 0.083]), and this increased over time (0.003 CrI=[0.001, 0.005]). This finding is consistent with the expected pattern of convention formation.

Table 2: Examples of cosine similarities between pairs of descriptions.

Expression 1	Expression 2	Sim
the red center	red middle	0.78
droopy iris flower	multi flowers	0.56
droopy iris flower	droopy	0.56
cluster	bundle	0.25
red middle	black background	0.25
droopy iris flower	the red center	0.09
droopy	bundle	0.03

We also examined how referring expressions related to the conventionalized names for other flowers. Utterances diverge from the conventionalized names of other flowers over time (trial: -0.003 CrI=[-0.003, -0.003], see Figure 5 lower panels). There was no effect of the same person saying the description and convention (0.001 CrI=[-0.005, 0.007]) and no interaction between same person and trial number (0 CrI=[0, 0.001]).

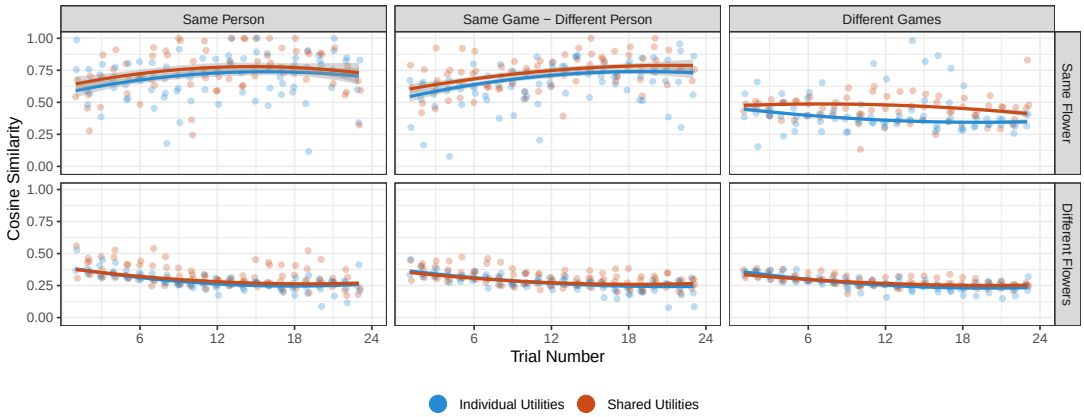

Figure 3: Cosine similarities between SBERT embeddings of utterances produced within 2 trials of each other. Utterances were paired with other utterances from the same speaker; from the same game, but a different speaker; or from a different game (vertical panels). Utterances were either in reference to the same flower or different flowers (horizontal panels).

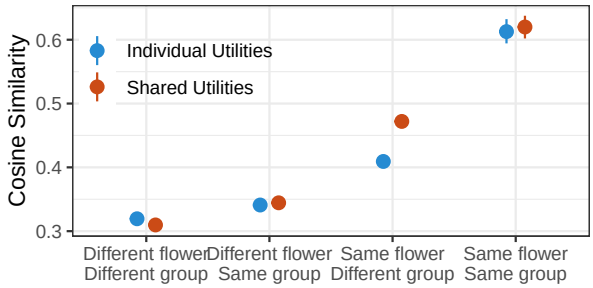

Figure 4: Cosine similarities between flower descriptions provided by participants after the game. Error bars are 95% CIs, but some are hidden by the point size.

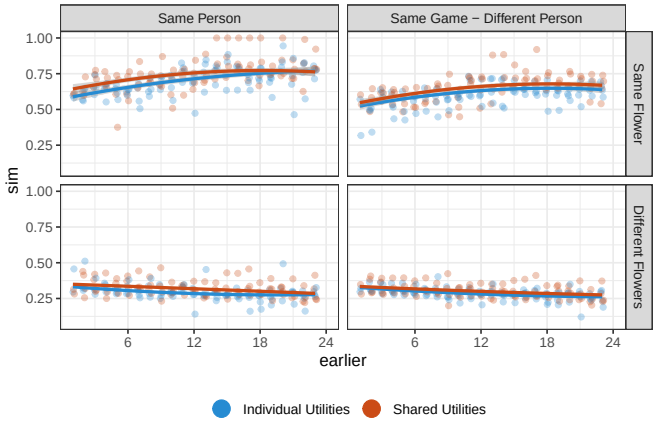

Figure 5: Cosine similarities between SBERT embeddings of utterances produced during a game and the post-game descriptions of the flowers.

Individual and shared utility conditions

The last question of interest is whether there were differences between the individual and shared utility conditions. Here we report the extent to which the shared utility condition differed from the individual utility condition. Overall, there was not a significant difference in the amount of referring language produced in the two conditions (0.076 CrI=[-0.661, 0.808]).

Within games Games in the shared utilities condition had non-significantly more similar descriptions of the same flower within games (0.056 CrI=[-0.023, 0.131]), but this did not interact with trial number (0 CrI=[-0.003, 0.003]) or whether the utterances were produced by the same participant (-0.014 CrI=[-0.077, 0.051]). Games in the shared utilities condition did have lower similarities between descriptions of different flowers within games (-0.015 CrI=[-0.024, -0.006]), but this was moderated by trial (0.002 CrI=[0.001, 0.002]) and by whether the utterances were produced by the same player (0.016 CrI=[0.001, 0.032]).

Comparing to the end conventions, games in the shared utility condition produced flower descriptions that were more similar to the convention for that flower (0.032 CrI=[0.007, 0.056]), but condition did not interact with trial number (0 CrI=[-0.002, 0.002]) and did not significantly interact with same participant versus different participant (0.032 CrI=[-0.011, 0.075]). In the shared utility condition, descriptions were slightly more similar to the conventions for other flowers (0.016 CrI=[0.011, 0.021]), but condition did not interact with trial (0 CrI=[0, 0]). This similarity was even higher if the description and convention came from the same person (0.019 CrI=[0.009, 0.027]). As a whole, groups in the shared utility condition may have had more internal alignment on the mapping between flowers and names, but this pattern was inconsistent.

Between games Across different games, games in the shared utility condition tended to have more similar descriptions for the same flower (0.052 CrI=[0.034, 0.071]), and this increased over time (0.003 CrI=[0.002, 0.005]).

Comparison of end conventions Lastly, we looked at how the conventionalized names compared, both between members of the same group and across players in different groups

that saw the same color palette of flowers, and whether these comparisons differed by condition (Figure 4).

We treated different flowers described by different games as the baseline condition. For individual games, descriptions from groupmates were more similar (even for different flowers, 0.022 CrI=[0.016, 0.027]), descriptions of the same flower were more similar (even across games, 0.09 CrI=[0.083, 0.097]), and there was a large interaction effect, where descriptions of the same flower were very similar among groupmates (0.182 CrI=[0.164, 0.199]).

For shared utility games compared to individual utility games, descriptions were roughly equivalent at baseline (-0.01 CrI=[-0.012, -0.007]) and for descriptions from groupmates (0.013 CrI=[0.006, 0.02]). Descriptions were much more similar for the same flower (even across games, 0.072 CrI=[0.064, 0.081]), but this was balanced out by a negative interaction effect between same flower and same game (-0.068 CrI=[-0.092, -0.045]).

These findings are suggestive that in the shared utility condition, games differentiated less from each other, and their conventions may have been influence more by shared priors for how to describe the flowers more than the conventions in individual utility games were.

Discussion

Our main question was whether the reduction phenomena found in dyadic reference games would occur within a negotiation game. This analysis of the referring expressions in Mankewitz et al. (2021) examined three key questions: the reduction of referring expressions, the convergence of referring expressions to conventions, and the differences between the shared and individual utility conditions.

First, we saw reduction, as referring expressions tended to decrease in length over the course of the game. Second, we found the expected set of convergence and divergence patterns. References to the same flower within a game converged both to other descriptions used at that time point and to the post-game descriptions. This convergence was specific to flower and games, as references to different flowers within a game and references to the same flower across games both diverged.

These patterns of reduction and convergence confirm some generalizability of the reference game patterns to the freer-form and more naturalistic domain of a negotiation game. While still artificial, this game differed from classic reference games in that the set-up was symmetric, with all players having equivalent amounts of knowledge and authority, and that the need for reference was embedded within a more complex goal structure. This game also was atypical for reference games in that it had more than 2 players (although see Boyce et al., 2022; Yoon & Brown-Schmidt, 2019) and used photos as target images (although see Weber & Camerer, 2003).

Our third consideration was how the different utility conditions moderated the reduction and convergence effects. We found some modest condition differences within games. In addition, groups in the shared utility condition diverged less from each other. This pattern of condition differences is consistent with less idiosyncratic descriptions in the shared utility condition, and the modest magnitude of the effect is consistent with the relatively subtle nature of this manipulation, which did not have a large effect on choice behavior (Mankewitz et al., 2021).

The greater alignment between players (and between games) in the shared utility condition might be due to a perceived greater importance of communicating with each other and understanding what others say, possibly triggered by the cooperative goal. The similarity across games is particularly intriguing given Guilbeault, Baronchelli, & Centola (2021), which found that larger groups tend to all converge to similar conventions whereas smaller groups can support idiosyncratic conventions.

However, we are cautious not to over-interpret the condition differences. There were a limited number of groups in each condition, so these effects may have been driven or exacerbated by heterogeneity in which participants were in each group. These ideas will need to be tested further with larger samples and more different goal structures.

Our interpretation of results is limited by the dataset, which is small, especially compared to recent datasets of convention formation in reference games (Boyce et al., 2022; Hawkins et al., 2020). The dataset also has large item-level variation, as the images differ in how nameable the flowers are or how salient background features are.

This dataset is just the start of testing how well findings from reference games generalize to more complex conversational situations. One dimension that we were unable to explore here due to the design of the game is the idea of shared versus privileged knowledge (Heller, Grodner, & Tanenhaus, 2008; Keysar, Barr, Balin, & Brauner, 2000). In the flowers game, like in reference games, all players saw the same set of images (and knew they did). In real-world situations, interlocutors may have different sets of objects in their context and may not know what is in their communication partner's context. An important future direction is studying how repeated reference phenomena interact with different communication goals and contexts.

In conclusion, our analysis here showed that the phenomena of reduction and conventionalization occurs even in complex games with different incentives, asymmetric knowledge, and more open-ended negotiation and dialogue than that found in reference games. Additionally, we saw hints that different incentives may lightly moderate the formation of conventions, but that reduction and convention formation took place regardless of incentive condition. We hope future work will continue to develop our understanding of these phenomena in more natural settings.

References

Almaatouq, A., Becker, J., Houghton, J. P., Paton, N., Watts, D. J., & Whiting, M. E. (2020, December 30).

Empirica: A virtual lab for high-throughput macro-level experiments.

Boyce, V., Hawkins, R., Goodman, N., & Frank, M. (2022). Two's company but six is a crowd: Emergence of conventions in multiparty communication games. *CogSci*, *44*.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1482.

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

Guilbeault, D., Baronchelli, A., & Centola, D. (2021). Experimental evidence for scale-induced category convergence across populations. *Nature Communications*, *12*(1), 327.

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

Heller, D., Grodner, D., & Tanenhaus, M. K. (2008). The role of perspective in identifying domains of reference. *Cognition*, *108*(3), 831–836.

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.

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*.

Leung, A., Hawkins, R., & Yurovsky, D. (2020). Parents scaffold the formation of conversational pacts with their children. *CogSci*.

Mankewitz, J., Boyce, V., Waldon, B., Loukatou, G., Yu, D., Mu, J., ... Frank, M. C. (2021). Multi-party referential communication in complex strategic games. *ICLR 2021 Meaning in Context Workshop*.

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.

Reimers, N., & Gurevych, I. (2019, August 27). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. arXiv.

Stan Development Team. (2023). RStan: The R interface to Stan.

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

Yoon, S. O., & Brown-Schmidt, S. (2019). Audience Design in Multiparty Conversation. *Cognitive Science*, *43*(8).