Two’s company but six is a crowd: emergence of conventions in multiparty communication games

Anonymous CogSci submission

Abstract

In repeated reference games where a speaker describes the same set of images to a listener over a series of rounds, the number of words used decreases as the pair converge on ad-hoc names for the images. The dynamics of this efficient reference formation is well-studied in dyads; however much communication takes place in larger groups, which are rarely studied in this paradigm. The current work extends iterated reference games to groups of 2-6 people who rotate between speaker and listener roles in an online game with text-based communication. Across 98 games, 390 participants, and 116K words, we find high accuracy and patterns of reduction regardless of group size.

Keywords: Communication; Reference game; Convention; Reduction;

Introduction

Verbal communication is an integral part of our daily lives. We coordinate schedules with partners, socialize with friends over board games, learn and teach in seminar classes, and listen to podcasts. Our communicative environments range in size from one-on-one to small group to large group to broadcast, but the goal of efficient communication is held in common. One necessity for efficient communication is shared reference expressions; when referring to a thing or an idea, it needs some sort of name that the interlocutors will jointly understand. In many cases, there are widely shared conventionalized expressions, but in other cases, spontaneous ad-hoc expressions are needed.

The formation of these new reference expressions is well-studied in dyadic contexts; however, dynamics may be different in larger groups, which are less studied. Our current work builds on the dyadic reference game tradition by extending it to small groups.

Dyadic reference games

The typical paradigm for studying partner-specific referring expressions is an iterated reference game with asymmetric knowledge. That is, each round there is a speaker who knows what the target is and a listener who does not have this information, and tries to select the target from a pool of possibilities. In Clark & Wilkes-Gibbs (1986), each speaker described 12 tangram shapes in order, so their listener could correctly order the images. After receiving feedback, the pair repeated the task with the same images but a new order, for a total of 6 repeats. Crucially, Clark & Wilkes-Gibbs (1986) found that reference expressions condense over the course of repeated reference to the same image. Early descriptions are long and make reference to multiple features in the image, but in later iterations, definite shorthand names are used.

Recently, online participant recruitment and web-based experiments have made it possible to study this convergence in larger populations using a text-based communication interface. In Hawkins, Frank, & Goodman (2020), 83 pairs completed a cued version of the iterated reference experiment. On each trial, the speaker saw one image highlighted and described it to the listener who clicked on what they thought the target was. Both players received feedback before moving on to the next target image. All images were highlighted each block, for a total of 6 blocks. Speakers produced fewer words per image in later blocks than in earlier blocks, in line with results from face-to-face, oral paradigms.

While this reduction pattern is robust for dyads, less is known about the how utterances are adapted in larger groups. A couple of studies point to some potential difficulties in trying to communicate with multiple people at once.

Multi-party communication

Yoon & Brown-Schmidt (2019) had speakers complete a sorting task with some listeners, so that they would have a common ground of shared names for the images. Then in a test phase, the speaker described these images to a group of either all knowledgeable listeners from the sorting task, new listeners who had not done the sorting task, or a mix of knowledgeable and new listeners. Speakers produced longer utterances when any new listeners were present than with only experienced listeners. This might predict slower reduction in larger groups where there will inevitably be some variability in how people understand reference expressions. These studies included 3-hour experiments that were very time and labor intensive, but some of the questions about group dynamics may be addressable in online experiments taking advantage of natural variation in understanding without artificially inducing large knowledge differences.

It’s difficult to communicate with naive listeners, but it can be even harder to communicate with someone with entrenched preconceptions. Weber & Camerer (2003) induced these conceptual differences by having two pairs of people (each pair representing a “firm”) do an iterated reference game with the same set of pictures. After 20 rounds, there was a “merger” where the listener from one group joined the other group. The reference game continued with the speaker communicating to both their original listener and the new listener. After the merger, there was a jump in how long it took either listener to make a selection. Even after several more rounds, listeners were still not as fast as before the merger. With larger groups of people all speaking together, there’s a greater chance for different people to independently develop different conceptualizations of an image, and it may be difficult for them to understand each other or agree on a common term of reference.

One open question is what the relationship between number of people and the success of the communication. More people seems to make communication slower and more difficult, but by how much? Does adding a 4th person increase the difficulty more or less than adding a 3rd did?

Conventions across speakers

In general, listeners expect speakers to maintain conventions and stick to descriptions that were similar to successful descriptions. However, they are not suprised to hear different descriptions of a familiar object if it comes from a new speaker who just entered the room (Metzing & Brennan, 2003) or if a new listener is present(Yoon & Brown-Schmidt, 2014). It’s unclear how these expectations map onto a group of people rotating roles in the task who are all present the entire time. Do later speakers count as new, or are they expected to follow conventions they’ve already heard? Do additional non-new listeners license changes in descriptions?

Present work

To address some of these questions, we extend the dyadic repeated reference game paradigm of Hawkins et al. (2020) to games for 2-6 players who rotate between speaker and listener roles. We compare accuracy and reduction rates in groups of different sizes. This paradigm allows us to confirm that that these findings in dyads extend to larger groups.

1. Accuracy will increase across blocks.
2. Listeners will respond faster in later blocks.
3. Speakers will reduce their utterances (produce fewer words) in later blocks.

Additionally, we will be able to test for trends across group size on the following two questions.

4. Do smaller groups use shorter utterances and reduce faster than larger groups?
5. Is there more overlap in how speakers describe each image in smaller or larger groups?

Methods

Building on the methods of Hawkins et al. (2020), we used Empirica (Almaatouq et al., 2020) to create real-time multiplayer reference games. In each game, one of the players started as the speaker who saw an array of tangrams with one highlighted (Figure 1 and communicated which figure to click to the other players (listeners). After the speaker had identified each of the 12 images in turn, the speaker role rotated to another player and the process repeated with the same images. In total, there were 6 blocks, giving each player at least one chance to be the speaker. We recorded what participants said in the chat, as well as who selected what image and how long they took to make their selections. Code to run the experiment, as well as data and analysis code are available at TODO anonymous OSF clone. .

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study. Our preregistration is at `https://osf.io/cn9f4`. TODO check anonymity

Participants

Players	Partial	Complete
2	4	15
3	2	18
4	2	19
5	3	17
6	6	12

Table 1: Number of games run for each player count.

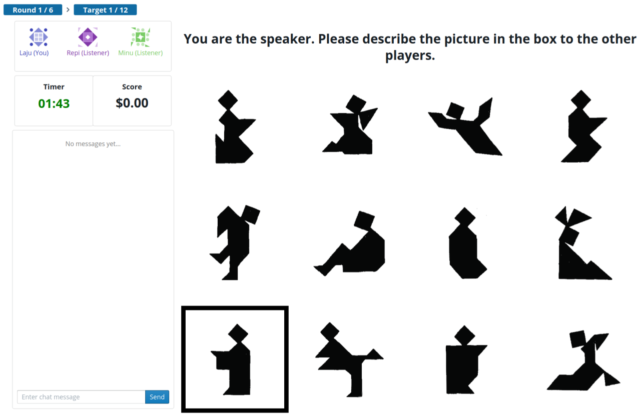
Participants were recruited using the Prolific platform between May and July 2021. We screened for participants who were fluent native English speakers. Participants were paid $7 for 2-player games, $8.50 for 3-player games, $10 for 4-player games, and $11 for 5- and 6-player games (with the intention of a $10 hourly rate), in addition to up to $2.88 in performance bonuses. A total of 390 people participated.

Materials

We used the 12 tangram images used by Hawkins et al. (2020) and Clark & Wilkes-Gibbs (1986) (see Figure 1). These images were displayed in a grid with order randomized for each participant (thus descriptions such as “top left” were ineffective as the image might be in a different place on the speaker’s and listeners’ screens). The same images were used every block.

Procedure

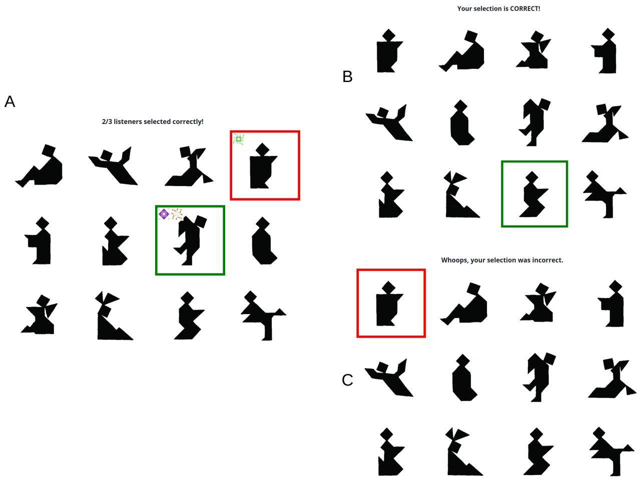
We implemented the experiment using Empirica, a Javascript-based platform for running real-time interactive experiments online (Almaatouq et al., 2020). From Prolific, participants were directed to our website where they navigated through a self-paced series of instruction pages explaining the game. Participants had to pass a quiz to be able to play the game. They were then directed to a “waiting room” screen until their partners were ready.



You are the speaker. Please describe the picture in the box to the other players.

No messages yet...

Figure 1: Screenshot of the speaker's view. Participants see all 12 tangram images.



2/3 listeners selected correctly!

Your selection is CORRECT!

Whoops, your selection was incorrect.

Figure 2: Screenshots of feedback for speakers and listners. Speakers (A) saw what figure each person chose, indicated by the matching icons. Listeners only learned if their selection was correct (B) or incorrect (C). Listeners were not shown what other listeners chose.

Once the game started, participants saw screens like Figure 1. Each trial, the speaker described the highlighted tangram image so that the listeners could identify and click it. All participants were free to use the chat box to communicate, but listeners could only click once the speaker had sent a message. Once a listener clicked, they could not change their selection.

Feedback Once all listeners had selected (or a 3-minute timer ran out), participants were given feedback (Figure). Listeners learned whether they individually had chosen correctly or not; listeners who were incorrect were not told the correct answer. The speaker saw which tangram each listener had selected. Listeners got 4 points for each correct answer; the speaker got points equal to the average of the listeners' points. These points translated into performance bonus at the end of the experiment.

In each block, each of the 12 tangrams was indicated to the speaker once. The same person was the speaker for an entire block, but participants rotated roles between blocks. Thus, over the course of the 6 blocks, participants were speakers 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.

After the game finished, participants were given a survey asking for optional demographic information and feedback on their experience with the game.

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 through the chat transcripts, tagging utterances that did not refer to the current tangram. These were primarily pleasantries ("Hello"), metacommentary about how well or fast the task was going, and confirmations or denials ("ok", "got it", "yes", "no"). We exclude 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.

Our intended sample size was 20 complete games in each group size, but we ended up with fewer as shown in Table 1. We excluded incomplete blocks from analyses, but included complete blocks from partial games. (Partial games occurred when a participant disconnected early, for example due to internet trouble.)

Results

Our first set of research questions were whether the classic findings of accuracy, speed, and reduction that are characteristic of two-player repeated reference games generalized to larger groups.

Accuracy and Speed

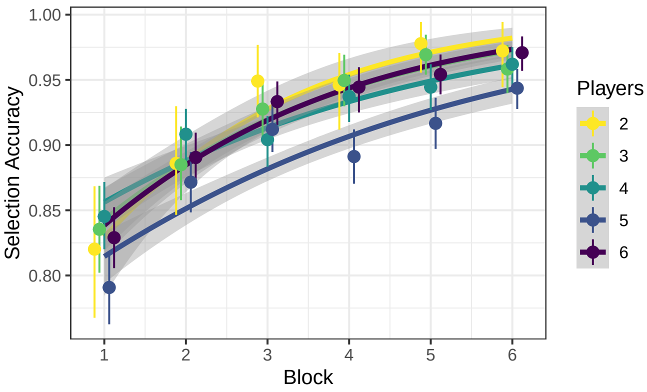


Figure 3: Player's accuracy at correctly selecting the target figure by block and group size. Accuracy increases across blocks.

Accuracy is high and increasing. Most individuals were accurate in their selections, with accuracy rising across blocks (Figure 3). In a logistic model of accuracy, participants are more accurate in later blocks (block: Est=0.38, CrI=[0.25, 0.5]), and there was no strong effect of group size on accuracy (numPlayers: Est=-0.02, CrI=[-0.08, 0.03]).

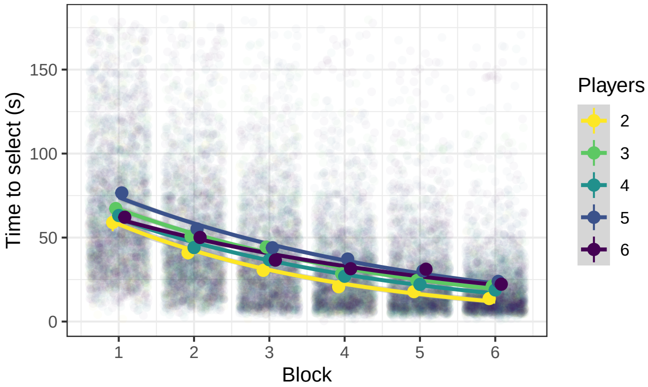


Figure 4: How long listeners took to select a figure in seconds by block and group size. Listeners selected images faster in later blocks. Only times for correct responses are shown.

Participants speed up in later blocks. Participants selected images faster in later blocks (Figure 4), although there was wide variability. In a linear model of selection time, participants got faster across blocks (block: Est=-10.03, CrI=[-11.03, -9.03]) and were slightly slower in larger games (numPlayers: Est=1.03, CrI=[0.4, 1.66]). This speed up is consistent with prior work by Weber & Camerer (2003) which used speed as the dependent measure. Wide variability in selection time meant that especially for larger groups, there was a wide spread in how long it took groups to complete the experiment.

Reduction

The key finding in dyadic reference games is that speakers produce shorter utterances as conventionalized names for the images arise. We replicate this finding in larger groups. Both speakers and listeners reduce the amount they say over the course of blocks.

Listeners rarely talk. Listeners often don't talk much, but are more likely to ask questions or make clarification in early blocks. In a linear regression for the number of words each listener said, there is an effect of block (block: Est=-0.48, CrI=[-0.79, -0.18]), but no clear effect of game size (numPlayers: Est=0.2, CrI=[-0.13, 0.51]).

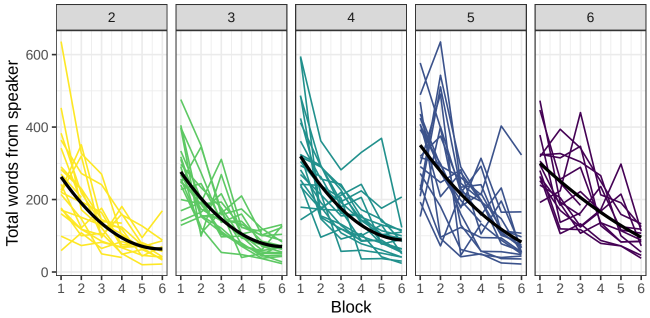


Figure 5: Number of words from speaker (total, across all 12 figures) in a block. Each colored line is one group, the overall trend is shown in black. Across group size, the number of words decreases as conventions emerge, but convention formation is not a smooth process, and there is variability between speakers.

Speakers utterances reduce in length. As shown in Figure 5, the number of words produced by speakers decreases over the course of rounds. This is true in aggregate, but also true for many groups. Nonetheless, in some groups, a later speaker may be more verbose than an earlier speaker. Speakers make longer utterances in early blocks that reduce to shorter utterances in later blocks. From a linear model, the effect of being one block later is block: Est=-3.35, CrI=[-4.58, -2.13].

Larger groups say more. One open question about larger groups was whether they would produce longer utterances or reduce more slowly than speakers in smaller groups. The overall effect of having more players in a group is numPlayers: Est=1.67, CrI=[0.68, 2.71] per additional player. There is no clear interaction between block and group size block:numPlayers: Est=-0.1, CrI=[-0.39, 0.18].

Development of conventions

One broad question is how speakers decide when to follow a previously established description and repeat it, perhaps in a reduced form, and when to use a different conceptualization.

When speakers don't know the convention In these games with limited feedback, listeners who got a tangram wrong don't have a way of knowing what the right answer was unless they ask for clarification in the chat. This means that if a speaker got a tangram wrong as a listener in the previous block, they may not know the conventional description that does with it, and thus are unlikely to follow it. If we assume that reduction is a sign of convention development, this predicts speakers should say more words when they got the tangram wrong the previous block, after controlling for other effects. This is borne out; speakers say more words for tangrams after they were incorrect (was_INcorrect: Est=3.07, CrI=[1.65, 4.5]).

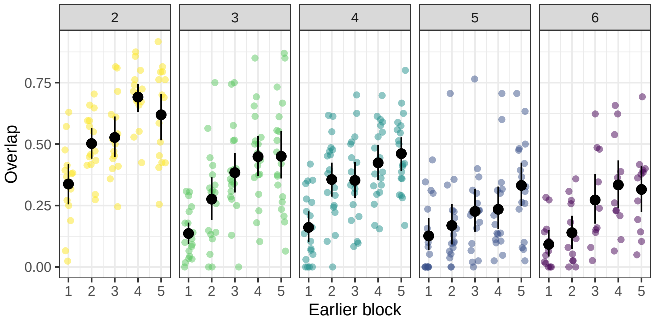


Figure 6: The fraction of content words used by the speaker in the last block that were used to describe the same tangram in an earlier block. Overlap is higher in smaller games.

Where do conventions come from? [TODO: check methods] Another angle to look at conventions is to take the speaker's utterances in the last block as the "convention", and look at how far back they started. To do this, we took the contentful words from the speakers last block utterances and looked at how many of them were used to describe that tangram in prior rounds. This let us calculate a fraction overlap between the last round utterance and earlier rounds, shown in Figure 6. In a linear model of the overlap between an earlier block and the last block, later blocks have more overlap block: Est=0.1, CrI=[0.08, 0.12]. Blocks with the same speaker as the last round have more overlap same_speaker: Est=0.08, CrI=[0.05, 0.1]; this is easiest to see in the peaks for rounds 2 and 4 in the 2 player games. Larger groups lead to less overlap between blocks numPlayers: Est=-0.07, CrI=[-0.09, -0.04], but there is no interaction between blocks and game size block:numPlayers: Est=0, CrI=[-0.01, 0]. Overall, this suggests that in smaller groups, conventions reduce and stabilize sooner, perhaps because fewer people need to implicitly agree on them. One potential confound is that in smaller games, players spend more time in the speaker role; however, there is still more overlap in smaller games even to blocks with a different speaker.

Examples While most groups did form conventions for most tangrams, it's illustrative to look at what happens when this doesn't happen. Table 2 shows the transcript of a 4-person groups for a specific figure where they described it geometrically every round, leading to long, and not very informative descriptions. Nearly all the figures have diamond heads, so this isn't a distinguishing feature, yet it is described. This illustrates the variability between groups, but also why conventions might be useful.

A different 4-person group had a member who during the first block shared the idea that the task would be easier if they explicitly gave "codenames" to the figures. The transcript for this group and one of the tangrams in shown in Table 3. Of note, multiple speakers forget the assigned codename, demonstrating that meta-knowledge doesn't always help. This group also describes the figure in relation to another already-named figured. Nonetheless, the group success-

Table 2: Excerpt from a group that did not reduce very much. The speaker for each round is marked with (S). Figure under discussion is row 3, column 3 in Figure 1.

Block	Person	Text
1	A(S)	Diamond on top. Body with no real arms or legs. The body is shaped like a boot with the diamond on top.
	C	Is the boot pointed left or right?
2	B(S)	diamond on top, large body beneath it. Left is a straight line all the way down, small variations on the right to the main body
3	C(S)	Diamond in center on top. Left side straight, right side carved out like a vase.
4	D(S)	Diamond head, flat topped body, straight on the left side with two triangles pointing out on the left
	D(S)	*on the right
5	A(S)	Diamond on top. Left side is straight, right side is obstructed, looks like a boot
	B	what do you mean by obstructed?
	A(S)	The left side of the body is right, right side has bents in it
6	B(S)	Diamond on top of a long large body/rectangle. Left side is complete, right side has bits missing

Table 3: Excerpt from a group that explicitly gave nicknames to the figures. The speaker for each round is marked with (S). Figure under discussion is row 1, column 4 in Figure 1.

Block	Person	Text
1	A(S)	[...] yes, the legs are like a zig zag
	C	CODE name ZIGZAG
	A(S)	There are no legs upwards
2	B(S)	okay so similar to begger guy but no foot pointing up
	B(S)	its like a zigzag
	B(S)	i forgot the code name
	D	zigzag yea
	A	The one standing with knees bent
	B(S)	yeah
	B(S)	standing
	C	Yeah zigzag
3	C(S)	The begger with no foot coming out from the left
	B	zigzag
	C(S)	zigzag it is
	C(S)	sorry i forgot
4	D(S)	zigzag
5	A(S)	zigzag
6	B(S)	beggar guy
	B(S)	zigzag

fully conventionalizes on a couple reduced names for this figure: "zigzag" and "beggar". This dual-naming of figures from multiple conceptual angles contributed by different speakers also occurs in other games.

Discussion

The overall patterns present in dyadic repeated reference games generalize to small groups of participants. Selection accuracy increases over rounds at the same time as listeners speed up their selections. Speakers reduce the length of their descriptive utterances as they conventionalize on concepts for each image.

Group size does make a difference, as larger groups talk more and are slower than smaller groups. There is also more diversity in how images are described within larger groups. This could just be from slower convergence to a conventions, or from parallel competing conceptualizations favored by different speakers.

The patterns of reduction are robust to speaker rotation and the lack of gold-standard feedback to listeners. Not only do speakers say less in later repetitions than they themselves said earlier, speakers later in the order say less than speakers earlier in the rotation. We provided less feedback than previous studies such as Hawkins et al. (2020). This low level of feedback means that there isn't a way for people to find out what was meant for utterances they initially did not understand outside of the verbal communication channel (or process of elimination). Similarly, speakers don't have direct access to how well their partners did in the previous block. Real-life communicative situations vary in what extra-textual feedback exists, but we do show that people can work around their initial confusion to eventually understand utterances, rather than just memorizing pairings after the first occurrence.

This is a rich data set that may yield more information about how humans adapt language quickly, and the dataset may be useful for training artificial agents to use and understand language more dynamically.

Acknowledgements

Place acknowledgments (including funding information) in a section at the end of the paper.

References

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]*. Retrieved from `http://arxiv.org/abs/2006.11398`

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

Hawkins, R. D., Frank, M. C., & Goodman, N. D. (2020). Characterizing the dynamics of learning in repeated reference games. *arXiv:1912.07199 [Cs]*. Retrieved from `http://arxiv.org/abs/1912.07199`

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. `http://doi.org/10.1016/S0749-596X(03)00028-7`

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. (2014). Adjusting conceptual pacts in three-party conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(4), 919–937. `http://doi.org/10.1037/a0036161`

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