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

Anonymous CogSci submission

Abstract

From classrooms to dinner parties, many of our everyday conversations take place in larger groups where speakers address multiple listeners at once. Such multiparty settings raise a number of challenges for classical theories of communication, which largely focus on dyadic interactions. In this study, we investigated how speakers adapt their referring expressions over time as a function of the feedback they receive from multiple parties. We collected a large corpus of multiparty repeated reference games (98 games, 390 participants, 116K words) where speakers designed referring expressions for groups of 1 to 5 listeners. Larger groups tended to use more words total and introduce more new words; nonetheless, most groups were able to converge to more efficient conventions regardless of the number of listeners.

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. Communicative environments range in size from one-on-one dialogue to broadcast communication to large groups, but the goal of efficient communication is shared across these. Shared referring expressions are a necessity for efficient communication; a thing or an idea needs some sort of name that the interlocutors will jointly understand. In many cases, there are widely shared conventionalized expressions for objects or ideas, but in other cases, spontaneous ad-hoc expressions must be invented.

The formation of these new reference expressions is well-studied in dyadic contexts and has been a case study for efficient communication more broadly. But these 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 larger groups.

Dyadic reference games

Clark & Wilkes-Gibbs (1986) established an experimental method for studying the emergence of new referring expressions that has now become standard. Two participants see the same set of tangram figures; the speaker describes each figure in turn so the listener can select the target from the set of figures. The speaker and listener repeat this process with the same images over a series of blocks. Early descriptions are long and make reference to multiple features in the figure, but in later iterations, shorthand conventional names for each figure tend to emerge; this shortening of utterances is called 'reduction'.

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 similar iterated reference experiment where they communicated via a chat box. Speakers reduced their utterances, producing fewer words per image in later blocks than in earlier blocks, in line with results from face-to-face, oral paradigms.

Multi-party communication

While in a dyad, the speaker can tailor their utterances to the one listener, in large groups, speakers must balance the competing needs of different listeners. One possible approach for speakers is to 'aim low' and produce utterances tailored to the least knowledgeable listener. For instance, in Yoon & Brown-Schmidt (2014), speakers develop conventions with one listener over the course of a reference game. If a new naive listener joined the group, speakers tended to use longer, more elaborated descriptions than if they continued talking with just the experienced listener.

Another strategy for speakers is to integrate across listeners and balance efficiency with informativeness by 'aiming in the middle'. In Yoon & Brown-Schmidt (2019), speakers play a reference game with a set of listeners, and then communicate to a mixed group of experienced and new listeners. In trials with 3 experienced listeners and 1 naive listener, speakers used shorter utterances and made fewer accommodations to naive listeners than in groups with a greater fraction of naive listeners. Both of these strategies predict that larger groups will be slower to converge than smaller groups, but differ in whether the the slowest listeners will be included.

Disagreements about how to conceptualize referents can also slow groups down. In Weber & Camerer (2003), pairs of participants played a reference game with the same image sets before a listener switched groups and joined a different pair. The addition of the new listener slowed both listeners down for multiple rounds. When a listener switched groups, they brought preconceptions about how to conceptualize the references, and there was conflict between how the new listener wanted images described and how the speaker was used to

describing the images. This predicts that, with more perspectives in play, larger groups may have more difficulty agreeing on common conceptualizations.

In general, listeners expect speakers to maintain conventions and stick to descriptions that were similar to successful descriptions. However, they are not surprised to hear different descriptions of a familiar object if it comes from a new speaker who just entered the room (Metzing & Brennan, 2003). It's unclear what this predicts about new speakers who were present as fellow listeners during prior blocks – will do listeners expect them to maintain the convention?

Work on multi-party communication has focused on the addition of a new person into a pair or group that had built up some shared representations. Our present work complements this by examining the effect of group size during the process of convention formation.

Present work

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 patterns 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 `https://osf.io/qdvbr/?view_only=47aebfde243f405e9c42a45cacb697d2`.

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study. Our preregistrations are at `https://osf.io/cn9f4/?view_only=7fdacd698b24465cb1a8699050af5bfc` and `https://osf.io/rpz67?view_only=5284203e2b644fc5ac39cf3e723b9a7e`.

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.

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. There was no signal to the speaker or other listeners about who had already made a selection.

Feedback Once all listeners had selected (or a 3-minute timer ran out), participants were given feedback (Figure 1). Listeners learned whether they individually had chosen correctly or not; listeners who were incorrect were not told the

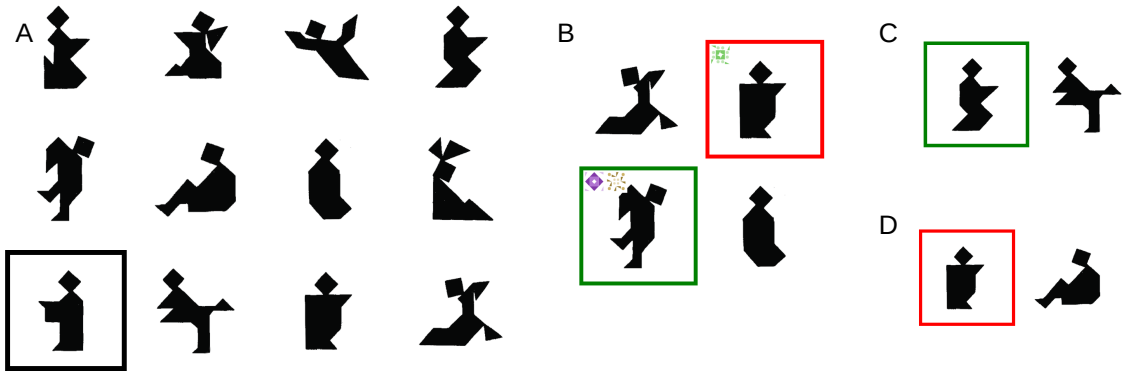

Figure 1: Speaker's view during selection phase (A). Participants see all 12 tangram images. During the feedback stage, Speakers (B) saw what figure each person chose, indicated by the matching icons. Listeners only learned if their selection was correct (C) or incorrect (D). Listeners were not shown what other listeners chose.

correct answer. The speaker saw which tangram each listener had selected, but listeners did not. In contrast with prior work (Hawkins et al., 2020) that told all listeners the right answer during the feedback stage, as we were concerned that listeners could learn arbitrary mappings just based on the pairing of what the speaker said and what turned out to be the right answer, without understanding the convention. Without this backchannel, listeners would need to learn the conventions just via the communication channel, which provides a stricter test of how well a group can converge on a convention.

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. Rotating the speaker was chosen to keep participants more equally engaged (the speaker role is more work), and to give a more robust test for reduction and convention.

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"), meta-commentary 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

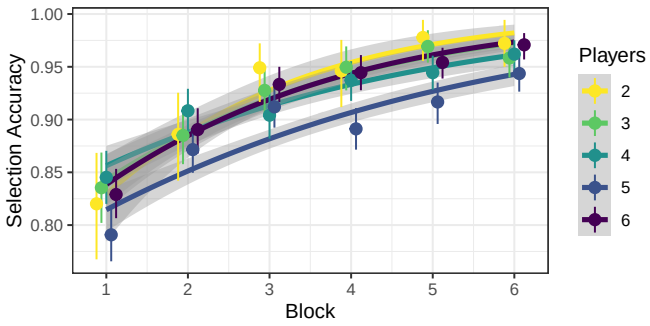

Figure 2: Player's accuracy at correctly selecting the target figure by block and group size. Accuracy increases across blocks.

Accuracy is high and increasing. Most individuals were accurate in their selections, with accuracy rising across blocks (Figure 2). In a logistic model of accuracy (correct.num$\sim$ block $\times$ numPlayers), 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]) or interaction between block and group size (block:numPlayers: Est=-0.01, CrI=[-0.04, 0.01]).

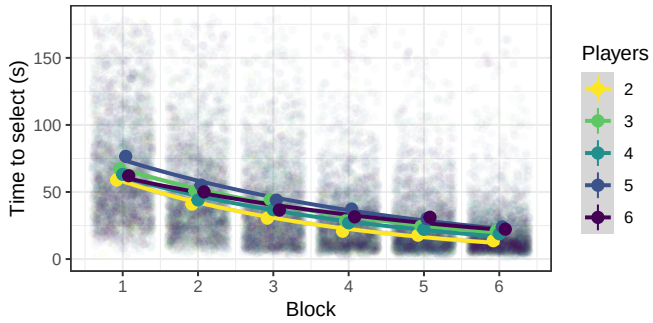

Figure 3: How long listeners took to select a figure in seconds by block and group size. Listeners selected images faster in later blocks. Only times for correct responses are shown.

Participants speed up in later blocks. Participants selected images faster in later blocks (Figure 3), although there was wide variability. In a linear model of selection time (time$\sim$ block $\times$ numPlayers), 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,[1] 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]).

Speakers utterances reduce in length. As shown in Figure 4, 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[2], the effect of being one block later is -3.35 (CrI=[-4.58, -2.13]).

[1] words$\sim$ block $\times$ numPlayers + (block|tangram) + (1|playerId) + (1|tangram_group) + (block|gameId)

[2] words$\sim$ block $\times$ numPlayers + (block|tangram) + (1|playerId) + (1|tangram_group) + (block|gameId)

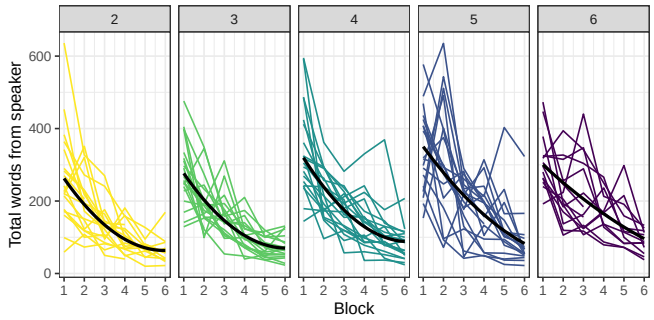

Figure 4: Number of words from speaker (total, across all 12 figures) in a block. Each colored line is one group, the overall trend is shown in black. Across group size, the number of words decreases as conventions emerge, but convention formation is not a smooth process, and there is variability between speakers.

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 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]). This is consistent with predictions from audience design that with more listeners to accommodate, the speaker may use multiple conceptualizations, either initially as a hedge or in response to listener clarifications.

Speaker experience does not fully explain group size effects. One potential concern is that group size correlates with whether the speaker has had the speaker role before (smaller groups repeat speakers more). To control for this, we add a variable coding for whether the speaker has been speaker in an earlier block[3]. Repeat speakers do use fewer words (speaker.repeat: Est=-8.55, CrI=[-10.41, -6.79]), but there are still effects of group size (numPlayers: Est=1.63, CrI=[0.58, 2.66]) and block (block: Est=-5.26, CrI=[-6.84, -3.69]). The effects of block and repeat speaker are subadditive (block:speaker.repeat: Est=3.2, CrI=[2.65, 3.78]), and there is minimal interaction between block and group size (block:numPlayers: Est=0.08, CrI=[-0.22, 0.41]).

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

[3] words$\sim$ block $\times$ numPlayers + block $\times$ speaker.repeat + (block|tangram) + (1|playerId) + (1|tangram_group) + (block|gameId)

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[4]; speakers say more words for tangrams after they were incorrect (was_INcorrect: Est=3.07, CrI=[1.65, 4.5]).

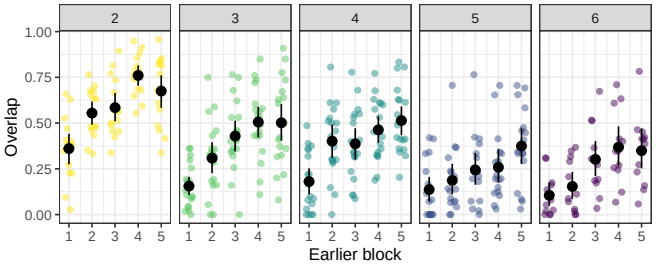

Figure 5: The fraction of content words used by the speaker in the final (6th) block that were used to describe the same tangram in each earlier block. Overlap is higher in smaller games.

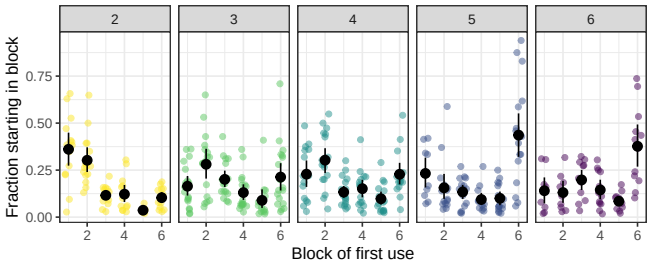

Figure 6: The fraction of content words used by the speaker in the final (6th) block that were first used by the speaker in a given block. More words are new in the 6th block in larger games; more words are first used in early blocks in 2-player games.

Where do conventions come from? 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 said by the speaker in the last block 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 5. In a linear model of the overlap between an earlier block and the last block[5], 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]). 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.

Another view of this is to look at when these words were first introduced by the speaker (Figure 6). A greater fraction of 6th block words are new in 5 or 6 player games compared with smaller games, whereas most words used in 2-player games originate in the 1st or 2nd blocks. Overall, this suggests that in smaller groups, conventions reduce and stabilize sooner, perhaps because fewer people need to implicitly agree on them.

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

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

[4] words$\sim$ block $\times$ numPlayers + block $\times$ was_INcorrect + (block|tangram) + (1|playerId) + (1|tangram_group) + (block|gameId)

[5] overlap$\sim$ block $\times$ numPlayers + same_speaker + (1|gameId) + (1|target)

Table 3: Excerpt from a group that explicitly gave nicknames to the figures. The speaker for each round is marked with (S). Tangram 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

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 successfully 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

We ran repeated reference games with groups of 2-6 players to see if three patterns from dyadic reference games generalized and test for two trends across group size, as layed out in the Introduction.

Consistent with dyadic games, listener's selection accuracy increased over blocks (prediction 1) at the same time as listeners sped up their selections (prediction 2). Crucially, speakers reduce the length of their descriptive utterances as they conventionalize on concepts for each image (prediction 3). Because speakers rotated, this reduction finding is robust: 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.

This reduction varies with group size; smaller groups use shorter utterances, but this does not significantly interact with block (question 4). The trajectory of reduction also depended on whether the current speaker correctly identified the tangram in the prior block and whether the current speaker was new to being speaker. This pattern is consistent with both the 'aim low' and 'aim middle' hypotheses.

Smaller groups showed more agreement in how each tangram was identified across blocks (question 5); the overlap between descriptions in the first 5 blocks to the final block was higher, and words in the final block tended to originate earlier. The greater diversity in how images are described in larger groups could be explained by slower convergence to a convention or parallel competing conceptualizations favored by different speakers.

We don't know what aspects of group size drive these trends as bigger groups differ from smaller groups in a number of ways. Larger groups more people for the speaker to communicate to, but also more people who might interrupt with questions, and more people who have opinions about what each image looks like.

Limitations

Group interactions are rich, and this experiment is necessarily a schematic simplification. Real-life situations vary widely in who the interlocuters are, their relationships, their goals, and their environment. Our participants were a convenience sample of Prolific workers who were strangers to each other; thus we miss richness that could come from prior relationships or shared community. Reference is only one goal out of many possible communicative goals, and the tangram images are artificial. We provided less feedback than previous studies such as Hawkins et al. (2020); this imitates situations where interlocuters can't show each other examples, but it's not representative of all communicative environments. Our text-based online paradigm meant that participant's individual identities were not especially salient. Communication takes place in a plethora of situations; our experiment provides some insight, but misses many complexities.

Future work

Iterations on this experimental paradigm could start to disentangle the mechanisms of group size and determine which design parameters are relevant to reduction. Luckily, with an online implementation, recruiting for and running experiments is feasible, and thus it will be possible to iterate on this experiment to determine how far the patterns generalize.

While much is left to be explored, this initial data set provides a rich corpus of how humans adapt language dynamically to communicate.

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