

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 have largely focused on dyadic interactions. In this study, we investigate 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. We find that speakers initially produce longer utterances when addressing more listeners, but that groups are able to converge to more efficient conventions regardless of the number of listeners. These findings contribute to understanding how shared viewpoints and concepts develop. TODO fix bigger picture statement

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.

Multi-party communication

The challenge of for speakers when there are multiple listeners is that they must choose utterances that will be effective to listeners who may have different background knowledge. One possibility is that speakers ‘aim low’, trying to be understood by all parties independently, leading to longer descriptions that focus on the quirks of the listeners who are struggling the most. For example, if a speaker is trying to

communicate with an experienced listener and a naive listener, they use longer, more elaborated descriptions than if they were only communicating with an experienced listener (Yoon & Brown-Schmidt, 2014). This suggests that communication will be slowed down when more listeners are present.

Another possibility is that speakers integrate across listeners and balance efficiency with informativeness by ‘aiming in the middle’. When some of the listeners are naive and need longer descriptions, speakers tailor utterances to them; however, speakers make fewer accommodations when there are 3 knowledgeable listeners and only 1 naive listener compared to when the fraction of naive listeners is higher (Yoon & Brown-Schmidt, 2019). This suggests that there may be non-linear relationships between group size and communication speed, as the speaker aims to be informative to most listeners, but without necessarily accommodating outlier listeners.

Communication can be slower and more tense when there are disagreements on how to conceptualize certain referents. When a speaker-listener pair was merged with an additional listener from another pair, the new listener had different expectations about what the key features of the referents were (Weber & Camerer, 2003). After the addition of the new listener, there was a jump in how long it took either listener to make a selection; this slowness persisted for several rounds. Weber & Camerer (2003) reported that some new listeners expressed frustration when speakers were describing images in terms of features other than the ones they were used to. The larger the group, the more difficult it may be for everyone to agree on a common conceptualization.

Work on multi-party communication strongly suggests that more people will make communication slower and more elaborated, but this work has focused on the addition of a new listener into a pair or group that had built up some shared representations. This leaves open the question of how group size influences the speed of representation formations to start.

Convention formation

Over the course of repeated reference to the same set of images with the same partner, people develop short-hand conventional names for each object. Early descriptions are long and make reference to multiple features in the image, but in later iterations, definite shorthand names are used (Clark & Wilkes-Gibbs, 1986). 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 produced fewer words per image in later blocks than in earlier blocks, in line with results from face-to-face, oral paradigms.

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) or if a new listener is present (Yoon & Brown-Schmidt, 2014). The reduction pattern may depend on the listener assuming that similar (but shorter) descriptions will correspond to the same image, and new descriptions will not correspond to images with other shorthand names. It's unclear how these expectations map onto groups where the listener and speaker roles rotate, but all participants are involved the entire time. Will later speakers be expected to maintain and develop previously established descriptions? Will the presence of other, possibly confused, 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 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 multi-player 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](https://osf.io/cn9f4) 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](https://osf.io/cn9f4) 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.

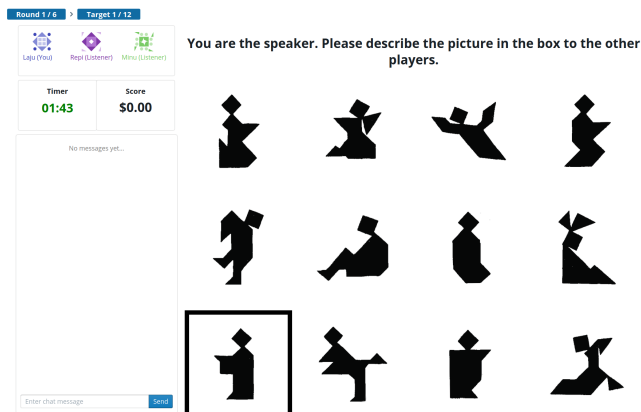


Figure 1: Screenshot of the speaker’s view. Participants see all 12 tangram images.

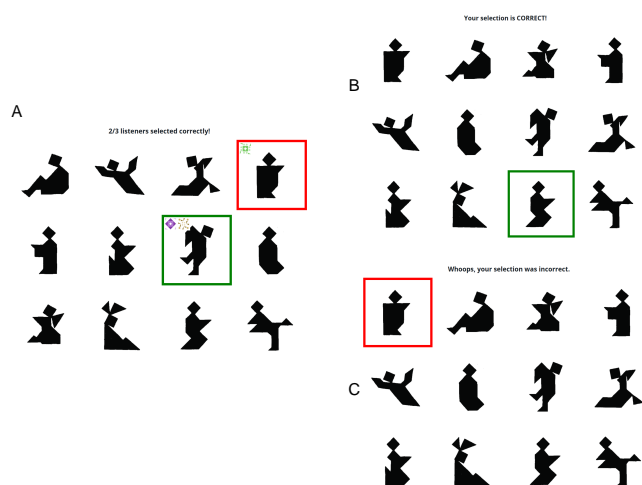


Figure 2: Screenshots of feedback for speakers and listeners. Speakers (A) saw what figure each person chose, indicated by the matching icons. Listeners only learned if their selection was correct (B) or incorrect (C). Listeners were not shown what other listeners chose.

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. 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). 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, 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

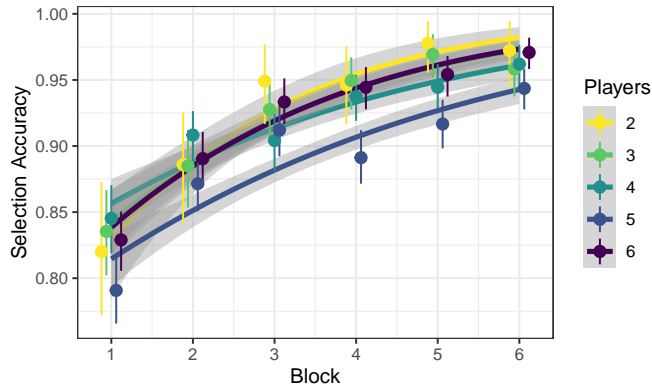


Figure 3: Player's accuracy at correctly selecting the target figure by block and group size. Accuracy increases across blocks.

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: $\text{Est}=0.38$, $\text{CrI}=[0.25, 0.5]$), and there was no strong effect of group size on accuracy (numPlayers: $\text{Est}=-0.02$, $\text{CrI}=[-0.08, 0.03]$) or interaction between block and group size ($r_{\text{stats}}(\text{acc}, 2)$).

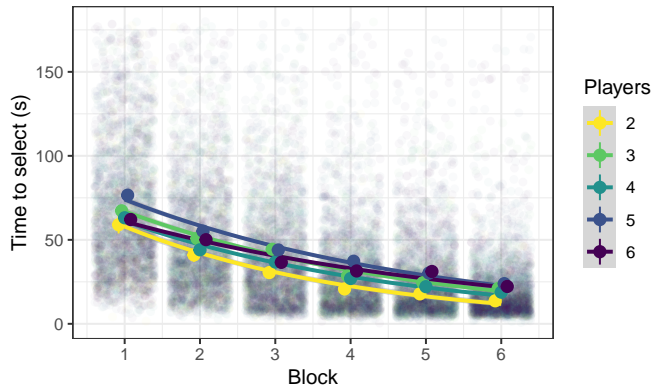


Figure 4: How long listeners took to select a figure in seconds by block and group size. Listeners selected images faster in later blocks. Only times for correct responses are shown.

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: $\text{Est}=-10.03$, $\text{CrI}=[-11.03, -9.03]$) and were slightly slower in larger games (numPlayers: $\text{Est}=1.03$, $\text{CrI}=[0.4, 1.66]$). 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: $\text{Est}=-0.48$, $\text{CrI}=[-0.79, -0.18]$), but no clear effect of game size (numPlayers: $\text{Est}=0.2$, $\text{CrI}=[-0.13, 0.51]$).

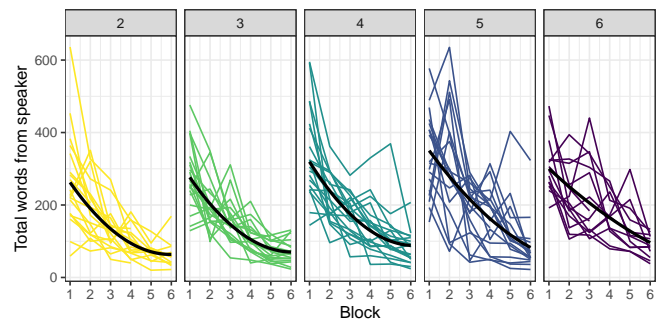


Figure 5: Number of words from speaker (total, across all 12 figures) in a block. Each colored line is one group, the overall trend is shown in black. Across group size, the number of words decreases as conventions emerge, but convention formation is not a smooth process, and there is variability between 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: $\text{Est}=-3.35$, $\text{CrI}=[-4.58, -2.13]$.

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: $\text{Est}=1.67$, $\text{CrI}=[0.68, 2.71]$ per additional player. There is no clear interaction between block and group size block:numPlayers: $\text{Est}=-0.1$, $\text{CrI}=[-0.39, 0.18]$. This is consistent with predictions from audience design that with more listeners to accommodate, the speaker may hedge with multiple conceptualizations. Larger groups also have more listeners who may ask clarifying questions that the speaker feels the need to respond to.

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]).

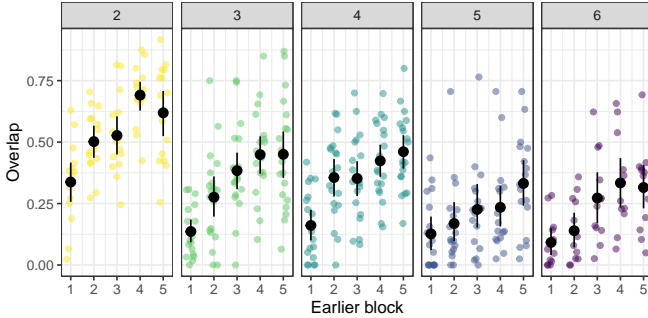


Figure 6: The fraction of content words used by the speaker in the final (6th) block that were used to describe the same tangram in each earlier block. Overlap is higher in smaller games.

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

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.
2	C	Is the boot pointed left or right?
	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

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 is 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

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
2	A(S)	There are no legs upwards
	B(S)	okay so similar to beggar 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
3	C	Yeah zigzag
	C(S)	The beggar with no foot coming out from the left
	B	zigzag
	C(S)	zigzag it is
4	D(S)	zigzag
	A(S)	zigzag
6	B(S)	beggar guy
	B(S)	zigzag

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.

Future work

There are a number of design parameters in this experiment that could effect the patterns of reduction. It remains to future work to vary these parameters and determine which are relevant. Luckily, with an online implementation, recruiting

for and running experiments is relatively easy, and thus it will be possible to iterate on this experiment to determine how far the patterns generalize.

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. These were fairly unfavorable settings where we limited the amount of information flow. Future experiments could see whether making more things public by making selections visible to all during the feedback stage changes the amount of reduction observed.

Additionally, while each participant had a nickname, color, and icon to identify them, individual identity was not especially salient. Future experiments could make the individual identity of each participant more salient by showing their responses in a consistent location on the screen, or allowing participants to get to know each other first. Different design choices in the interaction provide simulations for different contexts in which reference is established in normal life.

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](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>