TODO

Anonymous CogSci submission

Abstract

TODO

Keywords: Add your choice of indexing terms or keywords; kindly use a semi-colon; between each term.

introduction

Successful communication is grounded in a shared understanding of what utterances mean in the context they are produced in. 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 these ad-hoc referring expressions has been studied extensively in the context of dyadic reference games (Clark & Wilkes-Gibbs, 1986, ADD MORE CITES). 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 with a pair, as they tend to stick to the same description, but are ideosyncratic and differ between pairs.

This 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, etc [TODO]. The implicit theoretical claim is that these phenomena hold whenever people interact repeatedly in ways that require reference to some objects that do not have pre-conventionalized names that are adequate to distinguish them in context.

However, there are some ways in which this microcosm is not representative of typical language use. For one, in everyday communication, establishing reference rarely happens in isolation but is usually a subgoal in a larger conversational goal. The overall goal might be cooperative, such as in asking for a piece of kitchen equipment while cooking together, but it can also be adversarial, such as in a negotiation where the assets to be divied up must be referenced. In the reference game context, the goal is always cooperative, to match the images or get them in the same order.

The variation in goals can also lead to choices about when and what to refer to and how precise to be, depending on the costs of mistaken reference.

Do the results of dyadic reference games extend to more complex communicative interactions where goals are more complex and may not be aligned?

[possible worth acknowledging somewhere that sometimes esp. in legal things and philosophy, fighting over names is the argument, and we're going to ignore that]

[needs transition, but not sure how]

Mankewitz et al. (2021) tried to bridge the gap between datasets of negotiation over easily nameable objects, and the reference game literature where there is reference (but no negotiation) over hard to name objects. They created a 3 person game where players each select 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 between players. 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 saying that a player should refer to any specific flower. They also 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 hte game, possibly consistent with reduced referring expressions, but also consistent with developing a negotiation pattern.

Here we extracted the referring expressions from the chat logs of (**mankewitz?**) to look for reduction to partner-specific terms. We explored whether the phenomena found in classic dyadic iterated reference games extended to this more naturalistic domain, and whether they differed across the individual and shared incentive conditions. The key questions we addressed are:

1. Do the referring expressions get shorter over time?
2. Do referring expressions for the same image converge within a group?
3. Do referring expressions for different images diverge within a group?

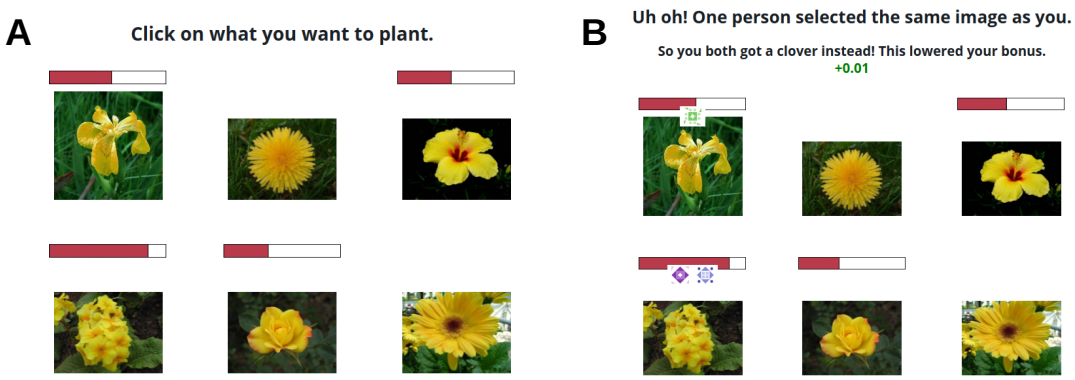

Figure 1: Player interface. Panel A shows the selection phase, where each participant sees 6 flowers, 4 with value bars. Panel B shows the feedback stage, after all players have selected. When multiple players select the same flower, they recieve a lower value rather than what is shown.

4. Do referring expressions for the same image diverge between groups?

   [TODO fix list of key points to match results.]

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.

Data Acquisition

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 four 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 selected a flower. 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). This incentivized participants to communicate about the flowers in order to coordinate on selecting differenct flowers. The rewards earned over the course of game translated in 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 earned 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. The use of 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 they had played with to their teammates. They were also asked how they would describe each of 4 images from a different color set than the one they had used.

Exclusions

Following Mankewitz et al. (2021) we only included games where players finished all 24 trials in the game. We only included games where participants had access tt the chat box (Mankewitz et al. (2021) also had a no-chat baseline condition). This left us with 18 games in the individual utilities condition and 21 in the shared utilities Condition.

Textual annotations

We annotated the chat transcripts to extract all referring expressions and identify which flower image each expression referred to. We also spellchecked and corrected the referring expressions. 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 annotators agreed on the exact reference expression for 105 (90%) of the cases. The second annotator 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 high level of interannotator agreement as an indication that the reference spans and targets were identified in a reliable way.

We extracted a total of 3395 referring expressions.

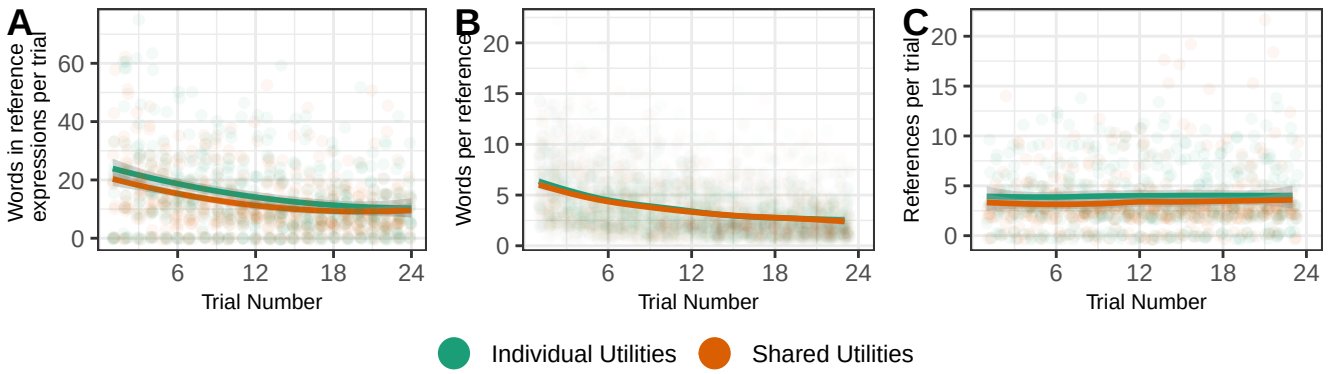

Figure 2: Amount of words produced in referring expressions across trial in games in both conditions. A: Total words of referring language per game per trial. B: Words in each reference expression by trial. C: Total number of reference expressions per game per trial.

SBERT embeddings

We embedded each of the extracted referring expressions using SBERT (Reimers & Gurevych, 2019), and used cosine distance between embedding pairs as a measure of similarity between the corresponding referring expressions.

For the pairwise comparisons of utterances during the game, we took the similarities for utterances that were within 2 trials of each other (ex. an utterance from trial 10 was compared with utterances from trials 8,9,10,11,12). Presentation of flowers was randomized, so different groups saw different flowers combinations on the same number trial. The width of the comparison window was a compromise between having enough pairs of utterances while still being able to treat them as coming from one time point in the game.

Results

We address each of the questions of interest listed at the end of the Introduction.

Reduction of referring expressions

The amount of referring language decreases 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 decreases across the game in both Shared and Individual utility conditions. The reduction in referring langauge was driven by shorter referring expressions later in the game (Figure 2B), while the total number of referring expressions per round remained constant (Figure 2C).

TODO include model!

Another coarse metric for how referring expressions are produced is how many different flowers are referred to each round. Each round of the game, there are 6 flowers visible on the screen, and participants want 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 round (Individual Utilities: 1.04 (sd: 0.92), Shared Utilities: 0.97 (sd: 0.78)).

The average pattern is consistent with each person saying what flower they plan on choosing. Some games talked less than this and risked collisions. Other games talked more as players queried the worths of various flowers or confirmed the plan of who would pick what.

? should there be a visual of these distributions ?

Convergence of referring expressions

TODO resummarize what cosine similarity is

[formatting comment – changes in cos sim tend to be small, so estimates and CredInts look kinda silly some of the time – could leave or increase sigfigs which is also silly]

Reference games have clear structures of repetition, as each image is usually the target once before any are again, and so it makes sense to treat each block as a time point. Because of the different structure of the game, trials are more continuous with each other. We smooth the random sampling and sparseness of flowers occuring the same trial by treating descriptions from within two trials as being at a comparable time in the game (so descriptions from round 10 are compared with descriptions from rounds 8, 9, 10, 11, and 12).

Within games One of the key claims is that within a partnership, the referring expressions should converge for the same image, but get more different for different images. Convergence would be increasing similarity (higher cosine similarity in later rounds), divergence would be decreasing similarity in later rounds).

If the dyadic pattern held, we expected similarity to increase over time for the same image with a group. In a model of similarity of utterances within flower and within group, we found that similarity increased over time (trial: 0.01 CrI=[0.01, 0.01], see Figure 3 lower middle and lower right panels). Utterances were more similar if they were produced by the same participant (0.06 CrI=[0.02, 0.1]), although this did not interact with trial number (0 CrI=[-0.01, 0]). Games in the shared utilities condition may have had higher similarities overall (0.06 CrI=[-0.02, 0.13]), but this did not interact with trial number (0 CrI=[0, 0]) or same par-

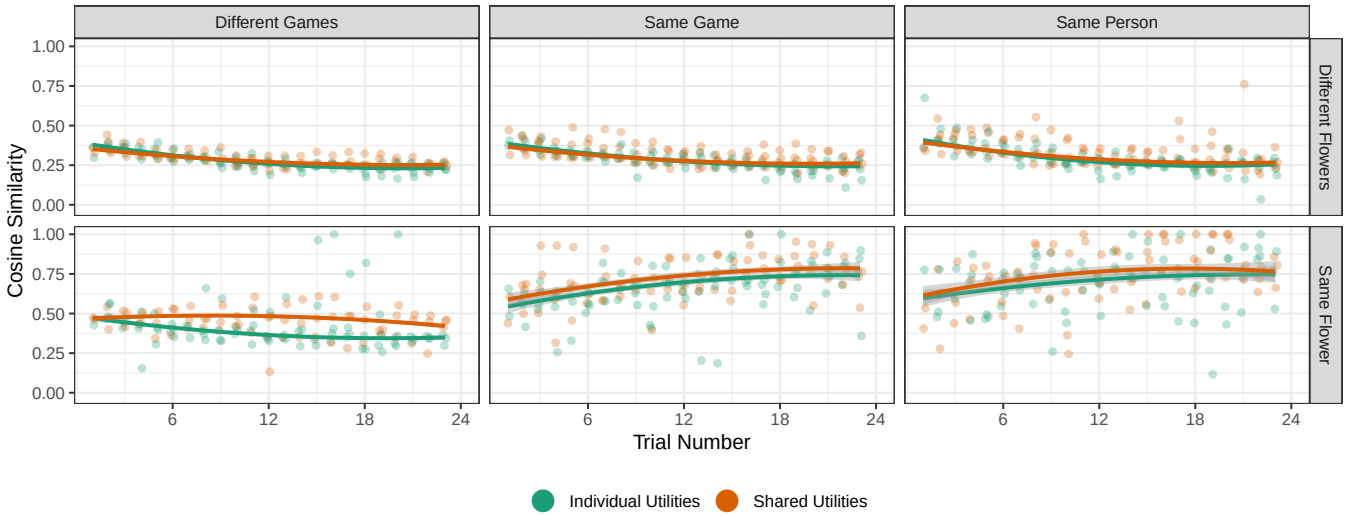

Figure 3: Cosine similarities between SBERT embeddings of utterances produced within 2 trials of each other. Utterances were paired with other utterances from different games; from the same game, but a different utterer speaker; or from the same speaker (vertical panels). Utterances were either in reference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

ticipant versus different participant (-0.01 CrI=[-0.08, 0.05]).

In contrast, we expected descriptions of different flowers to diverge within the same group. In a model of similarity between descriptions of different flowers within a group, similarity decreased over time (trial: -0.01 CrI=[-0.01, -0.01], see Figure 3 upper middle and upper right panels). Terms were slightly less similar in Shared Utility groups (-0.02 CrI=[-0.02, -0.01]), but this did not interact with trial (0 CrI=[0, 0]). There was not an effect of utterances being produced by the same person (0 CrI=[-0.01, 0.01]), or interactions between who said it and trial number (0 CrI=[0, 0]) or condition (0.02 CrI=[0, 0.03]).

The main results of convergence of descriptions for the same image and divergence between descriptions of different images matches the reference game phenomena.

Between games In reference games, conventions are partner-specific, so over time, descriptions become more particular to the group. If that held, we expected, utterances describing the same flower in different groups to have lower similarities later in the game.

In a model of utterance similarities for the same flower in different games, similiarities decreased slightly over time (trial: -0.01 CrI=[-0.01, 0], see Figure 3 lower left panel). There was a condition difference with shared utility conditions having more similar descriptions across games (0.05 CrI=[0.03, 0.07]); this did not interact with trial (0 CrI=[0, 0]).

TODO: what is our possibly explanation for this?!

This may be a fluke, as there are not a huge number of games in either condition, but if it's robust, it could be that the shared utility framing leads to less idiosyncratic descriptions and instead towards more universal priors. TODO this should

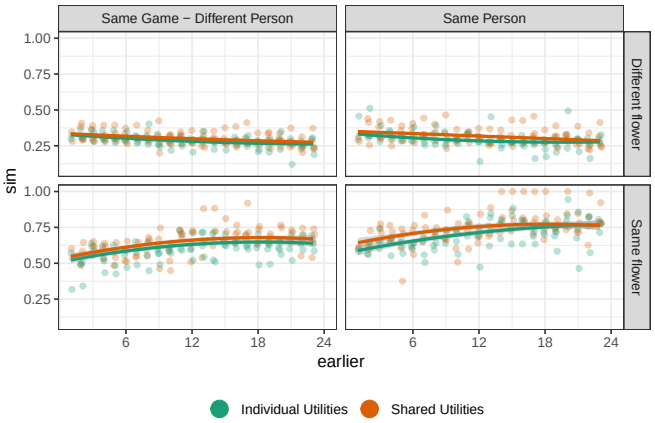

Figure 4: Cosine similarities between SBERT embeddings of utterances produced during a game and the post-game descriptions of the flowers.

be hedged *a lot* but we can also cite the study about large groupy things and the crab-like images

End expressions

game to end As a complementary way of looking at how referring expressions change over time as conventions are formed, we can take the post-game flower descriptions (each player was asked to provide what they would call each flower to their teammates) as the conventions, and look at how the in-game utterances converge to these conventions.

Descriptions become more similar to the conventions over time (trial: 0.01 CrI=[0, 0.01], see Figure 4 lower panels). Utterances were more similar to the convention given by the same person (versus from a groupmate) (0.05 CrI=[0.02, 0.08]), although this did not interact with trial number (0

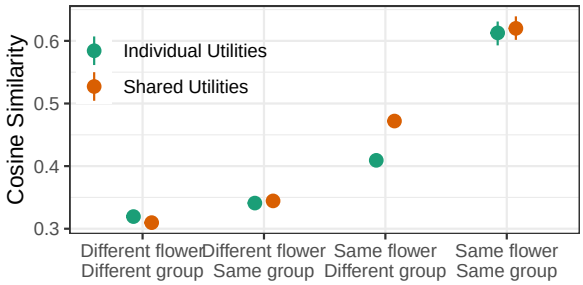

Figure 5: Cosine similarities between flower descriptions provided by participants after the game. Descriptions were more similar to each other if they came from the same group and were of the same flower.

CrI=[0, 0.01]). Games in the shared utilities condition had utterances that were more similar to their conventions (0.03 CrI=[0.01, 0.06]), but this did not interact with trial number (0 CrI=[0, 0]) and did not significantly interact with same participant versus different participant (0.03 CrI=[-0.01, 0.07]).

There was not a significant divergence from the conventions for other flowers over time (trial: 0 CrI=[0, 0], see Figure 4 lower panels). Terms are overall slightly less similar in Shared Utility groups (0.02 CrI=[0.01, 0.02]), but this does not interact with trial (0 CrI=[0, 0]). There is not an effect of being said by the same person (0 CrI=[-0.01, 0.01]), or interactions between who said it and trial number (0 CrI=[0, 0]). Similarity to a different convention was slightly higher if was said by the same person in a Shared Utility group (0.02 CrI=[0.01, 0.03]).

within end One last way of looking at convention formation is to look at how similar the end descriptions were to those of other players, either groupmates or people in different groups (but playing with the same color palette) for the same or different flowers. We expected that within a group, descriptions for the same flower would be much more similar, while descriptions of different flowers would be more different, compared to cross-group similarities.

As shown in Figure 5, compared to different flowers described by different games as a baseline, descriptions for the same game are more similar (even for different flowers, 0.02 CrI=[0.02, 0.03]), descriptions of the same flower are more similar (even across games, 0.09 CrI=[0.08, 0.1]), and there is a large interaction effect, where descriptions of the same flower are very similar among groupmates (0.18 CrI=[0.16, 0.2]). Compared with the individual utility condition, shared utility games had more similar descriptions of the same flower between games (0.07 CrI=[0.06, 0.08]). Other differences between the two conditions were small.

[probably need to talk about this result] [this whole section might be more interpretable if I centered Shared/indiv utilities in the model?]

Discussion

TODO!

Many of the key reduction findings from tangrams generalize to this situation. Specifically, we see utterance reduction over time and w/i group convergence for each image and divergence between images. This situation is different in that we have different stimuli (more natural) and the set up is collaborative and more free-form in what is talked about. These patterns hold for both the individual and group payoff structures.

One difference we see is that groups don't diverge (from each other). This may be dependent on stimulus properities (are there universal features of some of the images?) and group dynamics

mention the set of departures this made from classic reference game

Conclusion: The key reference game findings have some generalizability. Settings like this one may be useful for encouraging discussion of a set of images and setting up partial knowledge situations.

Acknowledgements

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

References

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

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. Retrieved from `http://www.speech.kth.se/~edlund/bielefeld/references/clark-and-wilkes-gibbs-1986.pdf`

Mankewitz, J., Boyce, V., Waldon, B., Loukatou, G., Yu, D., Mu, J., … Frank, M. C. (2021). *Multi-party referential communication in complex strategic games* (preprint). PsyArXiv. Retrieved from `https://osf.io/tfb3d`

Reimers, N., & Gurevych, I. (2019, August 27). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. arXiv. http://doi.org/10.48550/arXiv.1908.10084