TODO

Anonymous CogSci submission

Abstract

TODO

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

introduction

So people often care about reference or whatever. Referring expressions are crucial to interactive use of language so we can know what the interlocuter is talking about. Many goals of conversation rely on joint reference as a part of them – for instance in order to negotiate, even from an adversarial perspective, you still need to have terms that refer to the same things in order to do the negotiation. Joint reference is a cooperative act, needed for both cooperative and non-cooperative acts.

In some cases joint reference is easy b/c there are conventional meanings of words that refer to the objects. However, this isn't always the case, either because objects are novel and don't have names or because objects are similar to one another and so while they might be nameable out of context, in context there are not conventional names.

How people develop ad-hoc shared conventional referring expressions for objects without conventional names has primarily been studied in the realm of dyadic reference games. Two participants see a set of images, and the speaker sees a particular one highlighted as the one to pick out next (or they have them in the correct order and the listener needs to reorder theirs to match). This is a task explicitly about referring to the objects that is highly structured. There are assigned roles (one person has all the knowledge) and a clear indication of which one to talk about. Selecting or ordering objects is pretty simple.

This has been a very useful microcosm for studying referring expressions, expectations about conventions, and what happens with new listeners or speakers. The key phenomena observed across these experiments are that pairs develop partner specific referring expressions – the utterances produced by the speaker to refer to objects get shorter over repetition, the accuracy-speed trade off improves (? maybe don't mention this b/c it's not what we look at), and the terms that develop are partner specific – people converge on terms within partnerships, but differ between partnerships on the terms used.

Implicitly, these features are thought to be about how referring expressions develop across some wider range of situations where people interact repeatedly in ways that need reference to some objects that do not already have conventional names that are sufficient for reference. Basically it needs to be a situation where people don't at the outset have a high degree of agreement on how to refer to the thing.

Needs more transition here.

Mankewitz et al. (2021) tried to bridge the gap between datasets of negotiation (like let's make a deal?) where there is negotiation over easily nameable objects, and the reference game literature where there is reference (but no negotiation) to hard to name objects. They created a 3 person game where players had to select what flowers to grow from a set of 6 shown on the screen. The values of the flowers were partially hidden, and flowers were all the same color (but different shapes/species). Each player could select a flower, but it was only worth the shown value if one player selected it; if multiple did, they got a much lower reward. Flower images repeated across the game. (**mankewitz?**) intended this set-up so that players would need to refer the flowers in order to negotiate who would take what, but in a context where there wasn't a flower to talk about and there wasn't a designated speaker. 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). The paper 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.

This builds on the ref game literature in a few dimensions. The key difference is the more free-form format where reference is implicitly motivated by the task, rather than being an explicit part of the task, and the equal positions of the players (each has the same amount of knowledge) rather than an asymmetric teacher/student. Other differences include that the images are natural photos, which has been used in ref games (CITE that management paper), but is not the typical and the presence of more than two participants (which is done, but not modal, citations!).

Here we extract the referring expressions from the chat logs of (**mankewitz?**) in order to look at reduction as well as semantic trends. We see whether the phenomena found in classic dyadic iterated reference games extend to this more

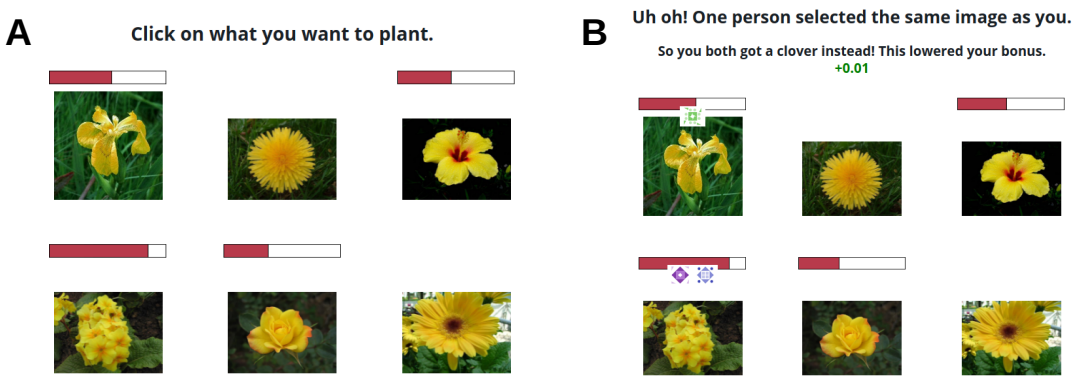

Figure 1: Player interface. Panel A shows the selection phase, where each participant sees 6 flowers, 4 with value bars. Panel B shows the feedback stage, after all players have selected. When multiple players select the same flower, they recieve a lower value rather than what is shown.

naturalistic domain, and whether they differ across the individual and shared incentive conditions.

TODO list of key points.

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.

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 assess the list of key points given at the beginning.

Reduction of referring expressions

Consistent with the results from dyadic iterated reference games, the amount of referring language decreases over the course of the game. As shown in Figure 2A, the number of words of referring language decreases across the game in both conditions. This is due to the individual referring expressions getting shorter (Figure 2B), while the total number of referring expressions remains that same (Figure 2C).

TODO include model!

Another coarse metric for how referring expressions are being produced is to look at how many different flowers are talked about each round. There are 6 visible on the screen, and participants are incentivized 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)).

This pattern is consistent with each person saying what flower they plan on choosing; some games talked less than this, and others talked more as players queried the worths of various flowers or confirmed the plan of who would pick what.

Convergence of referring expressions

Within games TODO we compare games within 2 rounds of each other (ex. trial 10 compared to 8, 9, 10, 11, 12)

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 similiarity (higher cosine similarity in later rounds), divergence would be decreasing similarity in later rounds).

For the same flower within the same game, similarity increased over rounds (trial: 0.01 CrI=[0.01, 0.01]). 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 participant versus different participant (-0.01 CrI=[-0.08, 0.05]).

Different flowers within the same game have diverging descriptions that get less similar over time (trial: -0.01 CrI=[-0.01, -0.01]). Terms are overall slightly less similar in Shared Utility groups (-0.02 CrI=[-0.02, -0.01]), 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]) or condition (0.02 CrI=[0, 0.03]).

Between games Utterances describing the same flower in different games diverged slightly (trial: -0.01 CrI=[-0.01, 0]). 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 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.

Descriptions become more similar to the conventions over time (trial: 0.01 CrI=[0, 0.01]). Utterances were more similar to the convention given by the same person (versu from a groupmate) (0.05 CrI=[0.02, 0.08]), although this did not interact with trial number (0 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]). 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 want to look for more similar with a group than between groups

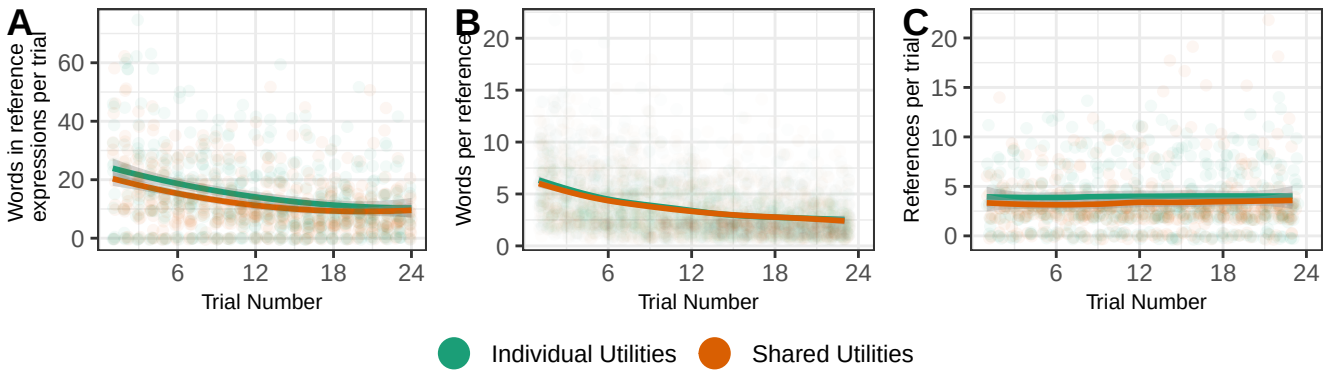

Figure 2: Amount of words produced in referring expressions across trial in games in both conditions. A: Total words of referring language per game per trial. B: Words in each reference expression by trial. C: Total number of reference expressions per game per trial.

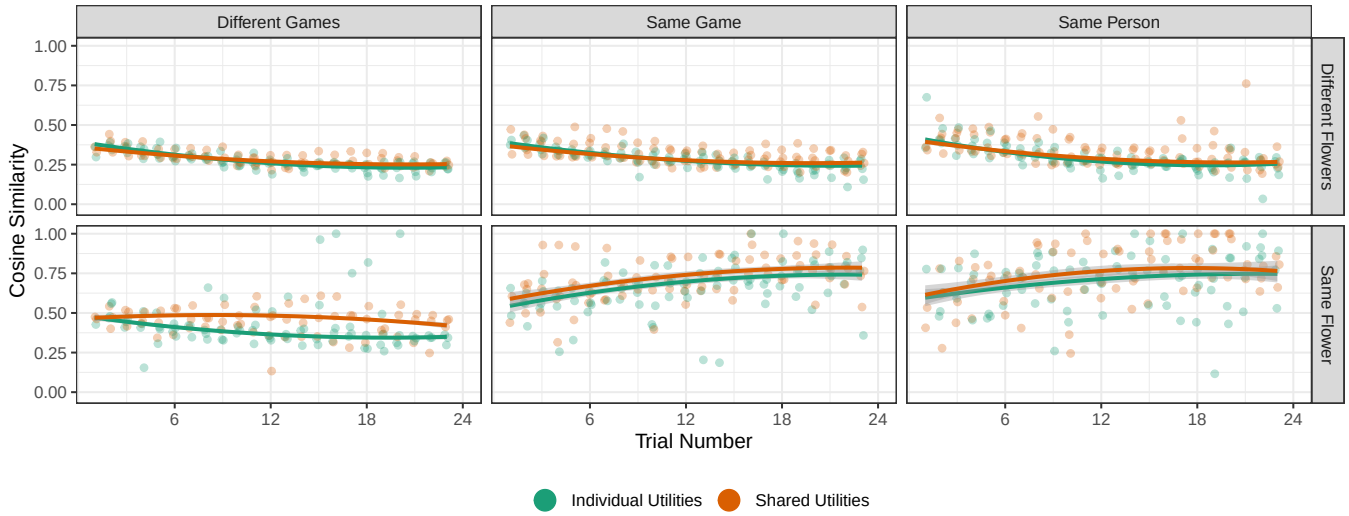

Figure 3: Cosine similarities between SBERT embeddings of utterances produced within 2 trials of each other. Utterances were paired with other utterances from different games; from the same game, but a different utterer speaker; or from the same speaker (vertical panels). Utterances were either in reference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

do we want to look at the "other" (flower not in game) at all? No?

Discussion

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

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`

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).

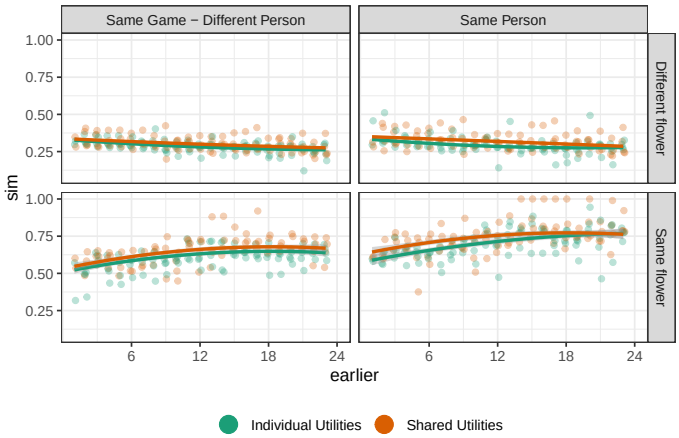

Figure 4: Cosine similarities between SBERT embeddings of utterances produced during a game and the post-game descriptions of the flowers.

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