TODO

Anonymous CogSci submission

Abstract

TODO abstract

Keywords: TODO keywords

Introduction

Conversation pacts and partner specificity are often studied by looking at how they are constructed; an additional perspective comes from how opaque or interpretable they are to outsiders who weren't part of the pact

By measuring opaqueness in different conditions related to how the pacts were formed or what the language looks like, can get another perspective on the process of pact formation

An empirical test of partner - specificity

Prior work to cover Summary of ref games & claims around them Hawkins et al. (2020) Clark & Wilkes-Gibbs (1986) etc

The side-participant / overhearer etc literature Wilkes-Gibbs & Clark (1992), & lit search for more

Judy's work Visual resemblance and interaction history jointly constrain pictorial meaning (**hawkinsb?**)

possible could also mention other times when naive comprehender has been used to better understand iteractive dialogues?

"Why use models" – can frame opaqueness as semantic distance between utterance and referent – models are an explicit test of this! "Naive comprehender" / "matcher"

ALvin gets to write computational intro here

Do we also want to motivate this from a computational angle? (i.e. trying to add pragmatics to models) TODO not me

back to Veronica

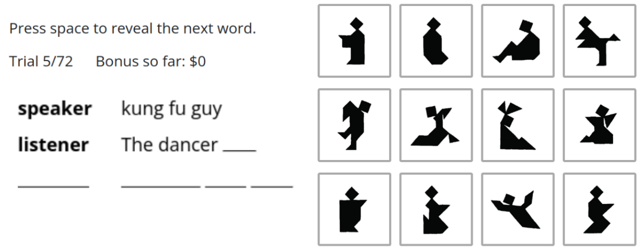
Key question: What properties of conversational pacts and the process of their formation make them more or less easy for an outsider to understand?

We use both human experiments and models to assess when and why expressions are opaque or understandable to outside observers.

Task Setup

Materials

We draw on the corpus of reference game transcripts and results from Boyce et al. (2024). There were all 6 round iterated reference games using the same 12 target images, but varied in how large the groups were (2-6 participants per group) and how "thick" the communication channel between group members was. For our human experiments, we sample different subsets of this corpus in different experiments. Within the samples, we avoided showing participants descriptions that had swear words or crude or sexual language. We use the entire corpus for our computational modelling component.



Press space to reveal the next word.

Trial 5/72 Bonus so far: $0

speaker kung fu guy

listener The dancer ____

Figure 1: Experimental Setup and Procedure.TODO REPLACE

Experimental procedure

We recruited participants from Prolific (TODO criteria). Participants were directed to the experiment, where it was explained that previously, other participants had described these shapes to one another. They would see a series of transcripts from the prior game, and their task was the guess what the intended target was. On each trial participants saw the full transcript from that trial, containing all the chat messages marked by whether they were from the speaker or a listener (TODO confirm the language used), except for lines that Boyce et al. (2024) had marked as not having any referential content. Participants selected the image they thought was the target from the tableau of 12. Participants received feedback on whether they were right or wrong on each trial. Except when the specific viewing order was part of the experimental manipulation, we randomized the order of trials, subject to the constraint that the same target could not repeat on adjacent trials.
The task was implemented in jsPsych. We paid participants $10 an hour plus a bonus of 5 cents per correct response.

Computational methods TODO V doesn't know how to write this QUESTION: do we focus on mlp pre- or post- calibration?

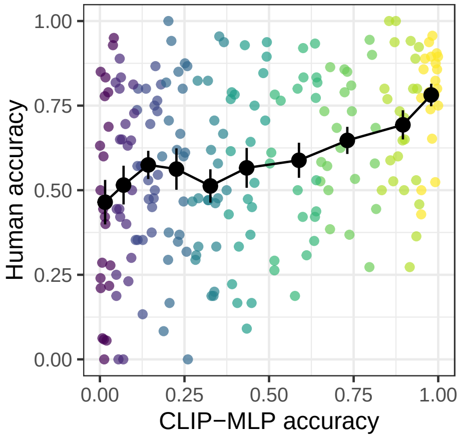


Figure 2: Correlation between human and CLIP-MLP accuracy across deciles of CLIP-MLP accuracy. Colored points are individual descriptions, black line is the boostrapped mean and 95% CI across descriptions for each decile.

Experiment 1

Our CLIP-MLP computational model was optimized for task accuracy. To validate whether this objective also results in human-like response patterns, we conducted a calibration experiment to determine if, for any given utterance, the probability that the model assigns to the target image is aligned with the probability that a naïve human matcher would choose the target image. We hypothesized that there would be a significant correlation between the target choice probability of the model and the target choice probability of naïve human matchers.

Methods

We first obtained target probabilities from our CLIP-MLP model for all utterances from Boyce et al. (2024). We then used stratified sampling to select 217 trials by dividing model-predicted probabilities into deciles and choosing approximately 22 utterances per decile, spanning the 12 different possible target images.

We recruited 61 participants who each saw 64 trials randomly sampled from the 217 tested trials. On average, each trial was seen by 18 participants. This experiment was preregistered at `https://osf.io/6pv5e`.

Results

We obtained human accuracies on each trial by dividing the number of participants who correctly selected the target image by the total number of participants who saw the trial, as shown in Figure 2. There was a small but significant positive correlation between model-predicted probabilities and human accuracies ($r = 0.3348057$, $p < .001$). This result suggests that model predictions were in fact calibrated to human response patterns, albeit not perfectly. It is theoretically possible to use these calibration results to project model predictions in order to better approximate human responses; we leave this approach for future work. Nonetheless, the observed positive correlation suggests that our computational model was a reasonable approximation of human accuracies, validating its use in subsequent experiments as a computational comparison.

Experiment 2

As a starting point for examining what makes referential expressions more or less opaque, we had people read the descriptions from the beginnings and ends of games. The idea of reduction and partner-specificity would suggest that the conventionalized, later round utterances would rely on the history of the game that naive matchers were not privy to, and thus that late round utterances would be more opaque and difficult to understand. On the other hand, describers gained practice over repetitions, so later round utterances might be better at clearly communicating the most visually salient features.

We included descriptions from games of different sizes and communication thicknesses. In terms of group conditions, based on the patterns of cross-game similarity in Boyce et al. (2024), we thought that smaller and thicker games were more likely to rapidly develop ideosyncratic conventions that would be more opaque than the less ideosyncratic conventions from larger groups with thinner communication channels.

Methods

Experiment 2a To establish a baseline of how well naive matchers could understand descriptions without context, we ran a 2x2 within subjects experiment. We drew the target transcripts from 2 and 6 player games from Experiment 1 of Boyce et al. (2024) and from the first and last (sixth) blocks of these games. These games had medium thick communication channels. We recruited 60 participants who each saw 60 trials (15 in each of the 4 conditions). Overall, participants saw 774 transcripts from 40 games. This experiment was preregistered at `https://osf.io/k45dr`.

Experiment 2b After observing limited condition differences in experiment 2a, we ran a second study drawing from the more extreme communication channel thicknesses of Experiment 3 in Boyce et al. (2024). Here, we used a 2x2x2 within subjects design, drawing our transcripts from the thick and thin, 2 and 6 person, 1st and 6th block utterances. In the thin condition, original matchers could only contribute to the chat by sending one of 4 emojis; as the emojis did not have referential content, we did not include them in the transcripts shown to naive matchers. For experiment 2b, we recruited 60 participants who each saw 64 trials (8 in each of the 8conditions). Overall, participants saw 2392 transcripts from 163 games. This experiment was pre-registered at `https://osf.io/rdp5k`.

Results

Our primary outcome was how accurate naive matchers would be at selecting the correct target, and how much this would vary depending on what game and what round the description came from.

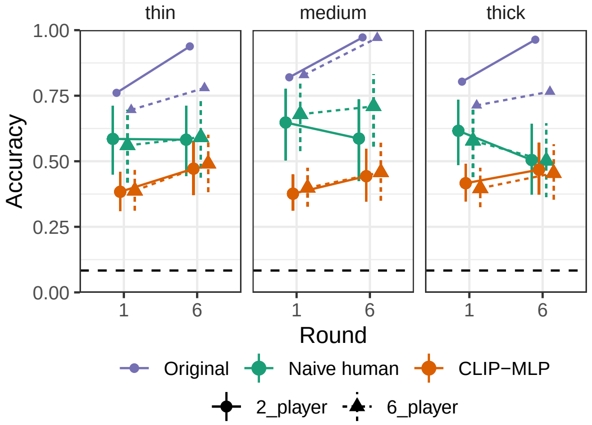


Figure 3: Accuracies for naive humans and the CLIP-MLP model for Experiment 2. Point estimates and 95% CrI are predictions from the fixed effects of logistic and beta regressions. Bootstrapped mean accuracy from the original matchers is included as a ceiling, and random chance as a baseline.

Experiment 2a For Experiment 2a, we ran a mixed effects logistic model of naive matcher accuracy: correct∼ group_size × round + trial_order + (group_size × round|correct_tangram) + (group_size × round + trial_order|workerid). Overall, naive matchers were right more often than not, which was far above the 1/12 expected by random chance3 (OR: 1.93 [1.05, 3.62]. As seen in Figure 3 middle panel, there were not large effects of condition. Participants tended to be less accurate at descriptions from the last round (OR of last round: 0.77 [0.53, 1.1]). There was not a clear effect of transcripts from 6-player games (OR: 1.15 [0.89, 1.47]), but there was an interaction between round and group size (OR: 1.49 [1.06, 2.1]). Later transcripts from larger games were easier to understand, but later transcripts from smaller games were easier to understand. Much of the variation in accuracy was instead driven by variation in the target images (OR of standard deviation of image distribution: 2.66 [1.88, 4.52]. Some images were much easier to identify as the target than others (Figure 4.

Experiment 2b For Experiment 2b we ran a similar mixed effects logistic model, consider the effects of group size, thickness, and round and their interactions. Overall, naive matchers were above 50% accuracy (OR: 1.81 [1.06, 3.08]). Similar to experiment 2a, there were not substantial effects of condition. Last round descriptions had slightly lower accuracy (OR of last round: 0.64 [0.47, 0.85]), but there was an interaction with thickness, where thin, last round were less opaque (OR: 1.55 [1.02, 2.33]).

Again some of the uncertainty in estimating the fixed effects was driven by the strong effects of target image (OR of SD of images: 2.25 [1.67, 3.59]).

Additional Predictors As additional post-hoc predictors, we also examined the predictive value of the accuracy of the

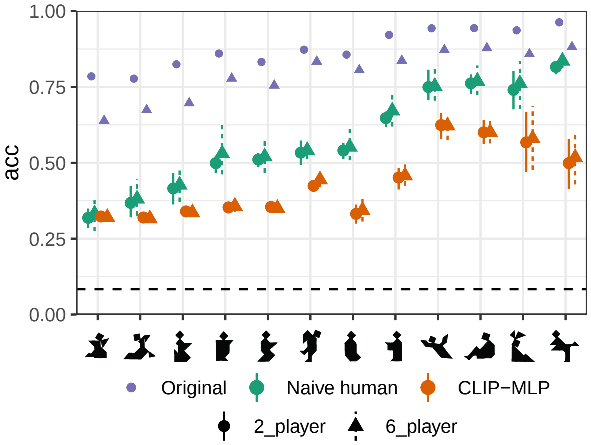


Figure 4: Accuracies for naive humans and the CLIP-MLP model for Experiment 2, split out by target image. Point estimates and 95% CI are predictions from the fixed effects and by-tangram random effects of logistic and beta regressions, bootstrapped across conditions. Bootstrapped mean accuracy from the original matchers is included as a ceiling, and random chance as a baseline.

original matchers from Boyce et al. (2024) and the the length of the description from the original describer. In both experiments, original accuracy was predictive of naive matcher accuracy (Expt 2a OR: 3.38 [2.46, 4.7], Expt 2b OR: 2.17 [1.7, 2.77]). The log number of words in the description was not predictive in Experiment 2a (OR: 1.05 [0.94, 1.17]), but longer descriptions were slightly beneficial in Experiment 2b (OR: 1.1 [1.01, 1.2]).

Model results

As a computational comparison, we looked at the CLIP-MLP model's performance on the same descriptions. We used the probability the model assigned as a measure of the model's accuracy, and fit a beta regression on the descriptions from Experiment 2: correct∼ group_size × thickness × round + (group_size × thickness × round|correct_tangram). The CLIP-MLP model was far above chance, but had lower accuracy than the human participants (OR: stats_text(acc_mod_mlp, 1) .

None of the fixed effects in the model were significant, and there was wide uncertainty for all of them. There is substantial by-tangram variation 1.58 [1.31, 2.15] and substantial by-tangram variation in the effect of later round 1.56 [1.29, 2.09].

As additional predictors, we checked the effect of original matcher accuracy and the length of the description. MLP-CLIP had higher accuracy when original matcher accuracy was higher (OR: 1.5 [1.33, 1.69]), and the model did better on shorter descriptions (OR for log words: 0.85 [0.82, 0.9]). Long descriptions may be further from the model's training distribution of image captions.

Interim Summary Overall, naive human matchers were fairly accurate overall, but less accurate than matchers in the original game. Perhaps surprisingly, this level of accuracy was fairly consistent across descriptions from different times in the game and different game conditions. The largest source of variability was from the target images; while there was some variabiliity in accuracy by images for the original matchers, there was substantially more variability for naive matchers.

Experiment 3

In Experiment 2, we saw that naive matchers could understand the descriptions fairly well, but had lower accuracy than the matchers in the original games. There are several differences between these two, including getting descriptions from a consistent group, getting descriptions in order, and being a present participant during the game. In Experiment 3, we focus on the role of context and group-specific interaction history to tease apart some of these differences.

Methods

Following CITE ROBERT AND JUDY, we compared naive matchers in "yoked" and "shuffled" conditions. In the "yoked" condition, naive matchers saw all the descriptions from a single game in the order they originally occurred. In the "shuffled" condition, naive matchers saw all the descriptions from a single game in a randomized order.

Because some descriptions are already pretty understandable in isolation, we wanted to focus on the role of context in games that showed strong group-specificity. We hand-picked 10 games from Boyce et al. (2024) on the basis of high original matcher accuracy, strong reduction in the length of utterances, and the use of idiosyncratic or non-modal referring expressions. Thus, the referring expressions were very understandable to groups who created them, but likely to be opaque out of context.

We recruited 196 participants (99 in the yoked condition and 97 in shuffled) who each saw all 72 trials of 1 of the 10 games. This experiment was pre-registered at `https://osf.io/zqwp5`. Participants read the transcripts in a modified self-paced reading procedure where they uncovered the text word by word (revealed words stayed visible); only after uncovering the entire transcript could participants select an image. We do not analyse the word-by-word reading data here.

Results

Our primary question of interest was how much having the conversation history would help make later round descriptions more understandable to participants in the yoked condition.

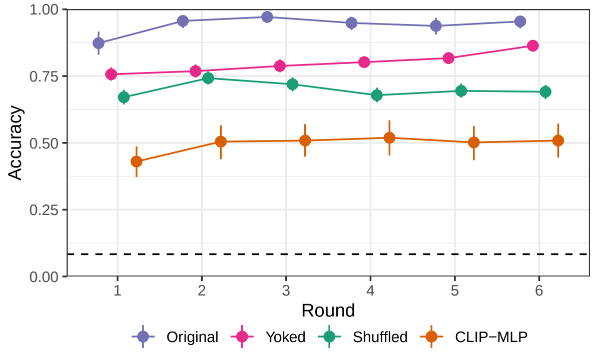


Figure 5: Accuracies for Experiment 3. Error bars are bootstrapped 95% CIs. TODO not using predictions because those fuzz out round to round differences.

We compared accuracy across the yoked and shuffled conditions with a logistic regression: correct∼ orig_repNum × condition + matcher_trialNum + (1|gameId) + (1|correct_tangram) + (1|workerid). The descriptions were more transparent when they were presented in a yoked order (OR: 2.2 [1.63, 3], Figure 5). In the shuffled condition, there was no main effect of round number (OR for one round later: 0.99 [0.95, 1.02]), but there was a marginal interaction where the benefit of the yoked condition decreased for later rounds (OR for one round later: 0.94 [0.89, 1]). This was offset by matchers in both conditions improving at the task over time (OR for one trial later in matcher viewing order: 1.02 [1.02, 1.02]). In the yoked condition round and trial number were aligned, so an improvement over time could be either from matcher practice or from descriptions being easier to understand. In the shuffled condition, matcher practice effects did not line up with position in the original game.

Comparing with the performance of the original matchers, we can separate out the benefits of seeing the descriptions in order versus being a live participant: correct∼ orig_repNum × order + orig_repNum × setting + matcher_trialNum + (1|gameId) + (1|correct_tangram) + (1|workerid). There is a benefit to seeing the items in order (OR: 2.24 [1.63, 3.04]) and a larger benefit to being a participant during the game in real-time (OR: 4.35 [2.77, 6.89]). The benefit of seeing the items in order wanes in later blocks (OR: 0.94 [0.89, 1]), but the benefit of being in the real-time game does not (OR: 1.06 [0.95, 1.18]). In all cases, there is a baseline improvement over trials (OR: 1.02 [1.02, 1.02]).

The accuracy of the CLIP-MLP model is worse than the shuffled human results, and does not show change across rounds (OR for one round later: 1.02 [0.97, 1.07]). The larger difference between naive human and CLIP-MLP accuracies in Experiment 3 than Experiment 2 could suggest that even the shuffled ordering still provides useful context that helps matchers understand the conventions. This history is not available to the CLIP-MLP model which sees every description as a one-shot task.

QUERY: we could run the model without matcher_trial_num which flips the direction of interactions because it centers later, and also you get to see the benefit of (matcher) experience purely in terms of the repNum. I think it's better the way it is, but raising it.

NOTE: not showing model estimates in Figure 5 b/c the fit isn't great, not sure why…

Discussion ?

Part of discussion that Alvin gets to write??

Discussion

Understanding varies much more based on item than on anything else; potentially due to priors or iconicity of image (? that might be beyond scope – how well does this match up with say diversity of descriptions)

Models do pretty well? IDK what our model take away is

Especially when there is strong or idiosyncratic reduction, context helps

role of context

limitations, incuding out of distribution for models

might want to address language comprehension v inference

References

Boyce, V., Hawkins, R., Goodman, N. D., & Frank, M. C. (2024). *Interaction structure constrains the emergence of conventions in group communication*.

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

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

Wilkes-Gibbs, D., & Clark, H. (1992). Coordinating beliefs in conversation. *Journal of Memory and Language*, 183–194.