Pragmatic adaptation in multi-party communication

Veronica Boyce

March 12, 2021

1 Introduction

Language is a crucial tool that we use to communicate in a multitude of contexts. A lot of the usefulness of language is driven by pragmatics, our ability to communicate using technically underspecified utterances by taking into account contextual factors, interlocutor knowledge, and conventional usage. This utility of language is reflective of social factors such as shared culture and the pressure language is under to be communicative. Verbal communication is a key part of human interaction: we use it to ask for objects, teach facts to others, and express our feelings. At a basic level, we use language to convey reference or information; the same principles of pragmatic language use also underpin more complex language use such as for instruction or deception.

1.1 Adaptation and Coordination

We tailor our communication depending on who we are talking to according to our prior interactions with them, what words they know, and what they know about the topic being discussed. This adaptation is dynamic; over the course of a conversation, shared terminology and shorthands naturally arise. These ad hoc descriptions may be intelligible only to the people who were part of the context in which they were established. If shared with more people and used more broadly, the terms might cement into a wider spread convention or jargon, but they might also fade away when the context ends.

This convergence to a shared shorthand, often called partner-specific adaptation, has been studied using reference games where one participant needs to use language to convey information only they have to the another participant who can do some task involving the designated object or image. With hard-to-describe objects people initially use long, hesitant descriptions and engage in back and forth clarification. With repitition, participants form a conceptual pact where they settle on what short name or phrase to use for each item. This adaptation is partner-specific, people know that these terms of reference are not widely understood, so if they need to communicate with others who are not part of the pact, they switch back to using longer, more elaborated descriptions. These phenomena have become a key model system for the study of communication more broadly.

1.2 Paper Outline

For my first year project, I propose an experiment looking at pact formation between multiparty groups. In this paper, I first discuss reference games in more detail including different variations used in studying partner coordination (Section 2). I then discuss a couple different theories of pragmatics and what predictions they make about iterated reference games (Section 3). Section 4 looks at contexts with more than two people, why these contexts are important, and what we know about pacts between more people. Section 5 summarizes the results in the context of the different theories. In Section 6, I outline my proposed experiment, and in Section 7, I mention some alternative interesting directions.

2 Iterated reference games

Reference games are a common psycholinguistic task where two people see a set of objects or images and one person tries to get the other person to pick out one of the objects with some term of reference. Depending on the context, this could be with unconstrained natural language, artificial language, or natural language with some constraints (only say one word). Often times, these are done as one-shot experiments where the trials are independent, but they can also be done as iterated reference games where the same objects and contexts repeat between the same two people. Over the course of iterations, pairs may establish implicit agreements about what terms to use for what objects. These conversation pacts form rapidly over the course of a few rounds, even between strangers.

These iterated reference games are used to study pragmatic communication because they are an easy way to elicit coordination and adaptation between speakers, which also occurs in more complex situations. In reference games, prior knowledge is more controlled; participants are strangers (although they still share the same cultural and language background) and the items are novel. Thus, reference games are a way of studying how these adaptations emerge, how they are negotiated, and what expectations people have around them.

Within this general class, there are variations of experimental paradigm. These variations provide converging evidence of partner-specific adaptation, but different variations involve properties of different real-life communication situations.

2.1 Task

In some tasks, the knowledge distribution is asymmetric, with one person (the director) knowing the desired goal, and the other participant (the matcher) following the directions to achieve that goal. In the **sorting** task, the director has a set of images in a set order and the matcher has the same set of images in a scrambled order [Clark and Wilkes-Gibbs, 1986]. The goal is to sequentially identify the images so the matcher can label or rearrange their images to match the director's order. While often done with images, variants such as arranging real toys and objects into a grid of cubby holes have similar properties [Metzing and Brennan, 2003]. In the complementary **picking** task, the director sees a set of images (often 4) with one highlighted. The matcher sees a shuffled order of the same images and the director needs to get the matcher to select the highlighted image. While both tasks can be used for both developing pacts and assessing accuracy, they are often paired, with several sorting rounds used to establish the references, and then picking rounds involving both familiar and novel images used to test accuracy and speed. I'd theorize that keeping the whole set of possible referents constant and visible may aid in picking out consistent distinguishing characteristics and achieving a stable one-to-one mapping, but I don't think the comparison has been done.

Another set of tasks are collaborative where knowledge distributions are complementary and both participants both have some of the knowledge and do some of the task. In the maze task used by [Garrod and Doherty, 1994], participants each navigate a grid, where some of the connections between squares are blocked and some are open. Each participant navigates their figure through the maze trying to reach a goal square, but there are also squares that flip which paths are open when the partner's figure enters the square. Thus, participants must coordinate to direct each other to the switch boxes as needed. Here the level of coordination is about how to refer to locations on the grid, whether by chess notation, cartesian coordinates, or path directions. Depending on the set-up of the grid, there may be more or less obvious ways of referring to locations.

Other collaborative tasks include ones where both participants independently select or arrange items, such as building Blokus animals in parallel [Ibarra and Tanenhaus, 2016]. On all these tasks, the goals of the participants are aligned; they both have the same shared goal.

2.2 Objects

For these tasks, the stimuli that will get coordinated reference need to be ambiguous enough that it isn't obvious what to call them, but concrete enough that some sort of system or analogy is used. One option is abstract shapes like tangrams where people will pattern match and see resemblances to humans or animals, but abstract enough that there won't be universally agreed on interpretations. This relies on humans ability to see patterns in many things. An alternative is busy images that share many features; [Weber and Camerer, 2003] used office scenes which tended to have people, desks, computer monitors and such in them, and so no picture to start with had a salient feature. Which features were important and were used thus differed between people. These sorts of busy scenes may make for better stimuli in human/computer interactions since identifying parts of an image may be easier for computer vision than the metaphoric use of tangrams. The main goal is that there isn't one dominant strategy, but that it's still possibly to converge to something.

2.3 Knowledge distribution

Conversational pacts are collaborative agreements between participants, but some of the set-ups use permanent asymmetric roles, which may weaken generalizations. For most sorting/picking tasks, the same person servers as director across all the rounds. This isn't a necessary condition of the design, for instance, [Weber and Camerer, 2003] have the participants switch roles each round in the training phase. Sometimes, the terms that get stuck are initially provided by the matcher, but with permanent roles, most of the language comes from the director. Thus, it isn't clear whether the same expectations around would also apply to the matcher if they became director, or if the dynamics and timeline of pact formation is different when the roles switch off. Other tasks are more collaborative, with each participant both having and needing some of the information [Ibarra and Tanenhaus, 2016, Garrod and Doherty, 1994].

2.4 Sample size

The scale at which these experiments are done varies greatly. [Clark and Wilkes-Gibbs, 1986] has a sample size of 8 dyads, and [Garrod and Doherty, 1994] compared a set of pairs to just one community, which makes it impossible to tell if the conventions the community settled on are due to sampling or the dynamics of the community. Some more recent studies have increased the sample size of groups; however, they are still limited to how many people can be brought into a lab and how readily audio can be transcribed. By running experiments online, using a chat system for communication, [Hawkins et al., 2020] was able to run experiments on hundreds of participants.

2.5 Medium

This different between text and spoken has some tradeoffs. Text is easier to analyse because it doesn't need to be transcribed, but it may have different dynamics than spoken speech. Oral communication conveys some more sources of information in particular around disfluencies and lengthened words (ex long e versus schwa as the vowel sound in the). As [Hawkins et al., 2020] notes, text-based communication leads to shorter utterances because there is less interruption. Text is much easier to study and the main patterns of results seem to hold up with text. Understanding the dynamics of text is also useful in and of itself, as it is becoming widespead as a medium of communication. Another potential reason for text-based research is the potential for eventually plugging in AI agents which is simpler in text than speech.

2.6 Measures of coordination

The key measure of coordination is how accurately the task is done. This is often taken for granted as performance is usually high, so properties of the communication that enable this performance are also examined. Some commonly used diagnostics of director speech are the length of the utterance in words (ex. [Yoon and Brown-Schmidt, 2019a]), the level of disfluency (ex. [Yoon and Brown-Schmidt, 2019a]), the amount of hedging in the utterances, and the number of elaborations in the utterances (ex. [Yoon and Brown-Schmidt, 2019a] One of the common measures of dialogue is the amount of turn taking, that is how often the matcher asks for clarification. In terms of performance, common metrics are how quickly and accurately the task is done. For instance [Weber and Camerer, 2003] explicitly rewarsd participants based on their speed at the sorting task. On a more semantic level, [Garrod and Doherty, 1994] looks at what sort of reference frameworks participants use (on the maze, counting lines, or boxes, or using a chess-like coordinate system) and when this changes over the course of a pact.

Because it provides a set of more formal techniques that I will be copying in my project, I go through in more depth the analyses done by [Hawkins et al., 2020]. They check that the length of director utterances in words do decrease over the course of the experiment. Going beyond that, they use SpaCy to provide part of speech tags and examine which words are dropping out and which are being kept, and find that nouns are often kept in. Looking at commonly dropped n-grams, they find evidence that hedges such as ‘looks like a’ are commonly dropped. By checking with a dependency parser, they also find that words tend to drop out as whole phrases. This decrease in utterance length is greater for tangrams that were correctly selected in the previous round, but it is tangram-specific: there isn’t an effect on the immediate next tangram (different tangram), only on the next occurrence of the same tangram.

In terms of the semantics, words that are initially used for fewer tangrams (that is, are more distinctive) are more likely to be preserved. They also measure the semantic idiosyncracies using GloVe embeddings averaged across all the content words in the sentence. In early rounds, embeddings for one tangram across pairs are similar, but they diverge as specific aspects are kept as salient by different pairs. Looking only at speaker utterances, the utterances within a pair are more similar than between pairs and seem to converge over repetitions as they become more similar and diverge from those of other pairs. Matchers send more messages in early blocks compared to later blocks, and these utterances are often questions.

3 Theory

There are a few prominent theories of pragmatics that can be applied to these reference games. All of these theories agree that the conceptual pacts form as a result of the social interaction between people. However there is disagreement on how to formalize theories and to what extent to focus on design choices the speaker reasons about versus bottom-up processes of alignment that occur unconsciously.

3.1 Qualitative descriptions

The starting point for pragmatics is the Gricean maxims that qualitatively describe some considerations for speakers to communicate effectively. In this descriptive vein, [**?**] lists a few factors that affect naming. The more an object is referred to by a name and the more recently that name has been used for that object the more likely it is to be used again (this is claimed to be partner-agnostic and occur even if settings without any feedback from the partner). When there is a specific partner, additional ‘grounding’ factors also contribute; a director may over a hedged provisional term and if this is accepted by the listener, it can be used more in shorter form and without the hedges, but only with that partner. These descriptions hold true experimentally, but raise the question of what sorts of processes could account for them and predict how they trade off with each other.

3.2 RSA

One computational formalism for pragmatic reasoning is the Rational Speech Acts (RSA) theory [Frank and Goodman, 2012] Building on game-theoretic models of human social communication, this theory posits that speakers and listeners iteratively use theory-of-mind to model the other's perspective and choose and interpret utterances accordingly. In many cases, this involves some counter-factual reasoning about alternatives. The RSA framework is usually used for non-iterated reference games and other situations in which meanings are invariant over the course of the game; however, the basic framework is flexible and can be extended to account for different situations, including those where meaning changes over time, such as word-learning or pact formation [Goodman and Frank, 2016].

One advantage of the RSA model over verbal models is it provides a way to make clear predictions about how different goals will trade off against each other – when does the value of recency override the value of quantity? Traditional qualitative descriptions don't answer this. In RSA, all of these goals are expressed in a utility theoretic way, and then the optimization process can weigh the different pressures against each other.

One downside of RSA is that it is a computational level model that does not specify how the iterated reasoning process might actually play out when it is approximated in a computationally tractable way. For instance, it seems like speakers sometimes overspecify rather than compute reasoning when it is ”cheaper” to do this, although it could also be that it's easier to overspecify than to risk that the listener might not compute the implicatures [Baumann et al., 2014]. The additional cost factors could be incorporated into a model, and in modelling iterated reference games, it may make sense to account for the speaker and listener figuring out whether the other participant is likely to overspecify or cooperate on pragmatic language use. These sorts of search costs could be incorporated into a model which might then be able to explain sub-optimal references, which would be important to accounting for references in a large search space, such as that presented when describing tangrams.

RSA mostly hasn't been used to model multi-party situations, but one simple RSA model of teaching is presented in [Frank and Liu, 2018]. This simulates a teacher who is trying to communicate the weight of a biased coin by showing the results of some flips to the students. Students vary in the prior beliefs about the weight of the coin, and update in a Bayesian fashion. At each time step, teachers can either give a lesson (show a flip) or give a test (get an example flip from each student). When students are assessed first and then grouped into small classes, performance improves because the students have more common ground that the teacher has access to, and thus can tailor the demonstrations to the students priors. How much assessment is useful depends on how extreme the target concept is. The same conflicts about who to teach to will arise in multi-player iterated games, where matchers may vary in what conceptual pacts they endorse, and directors need to somehow tailor utterances to a heterogeneous group.

3.3 Interactive Alignment Account

The Interactive Alignment Account focuses on the processing mechanisms of dialogue. Under this theory, people follow conversational precedents on many levels, including word choice and syntax (priming) which can lead to these precedents becoming conventionalized [Garrod and Doherty, 1994]. Rather than talk about conceptual pacts, this theory talks about alignment and the development of shared conceptual representations of a situation (such as a maze) [Pickering and Garrod, 2004]. This shared representation about the situation (what in other theories might be called common ground) is necessary for communication, even if the conversation goals are not aligned (i.e. one person wishes to deceive). In the Interactive Alignment Account, this shared basis is called the 'implicit common ground' and is developed without any explicit reasoning.

Under this theory, explicitly modelling other minds is treated as a difficult process that is only used as a last resort, but is not the default pattern[Garrod and Pickering, 2009]. It also treats conversation as a form of joint action; on some levels people are explicitly trying to coordinate (such as when people move furniture together), but on other levels such as priming, the alignment is unintentional (such as when people end up rocking in sync without intending to) [Garrod and Pickering, 2009]. They claim that these unintentional synchronies should be thought of as unintentional joint action, and that through relations between different levels of language production these low level alignments lead to shared representations. That is, more lexical overlap leads to stronger syntactic priming and the syntax and semantics of language have connections to the mental representations [Garrod and Pickering, 2009].

3.4 Audience design

Another theory about how speakers choose their utterances is audience design, the idea that speakers take listener knowledge and perspective into account when deciding what to say [Brown-Schmidt et al., 2015]. This makes similar predictions to RSA with low recursion depth, but has fleshed out and applied to multi-party communication more. When a speaker is talking to one other person, the relevant common ground is clear. When there are more people, there is no longer just one relevant common ground; some information may be shared by all participants, but some may be shared by the speaker and only some other participants.

One possibility is that speakers may ‘aim low’, tailoring utterances to the least informed listener and not relying on any incompletely shared common ground [Yoon and Brown-Schmidt, 2019a]. Another question of audience design is how finely people track who knows what – do they keep track if individual-level knowledge or just of an overall average idea of what is known [Yoon and Brown-Schmidt, 2019a]?

3.5 Relation between theories

All of these theories make the prediction that people should do well at the basic iterated reference game and jointly converge to an understanding. Theories disagree about how this happens, and thus make different predictions about what will happen in variant situations.

One big difference between RSA and the Interactive Alignment account is the relative role for social reasoning versus dialogue. The Interactive Alignment account does not make explicit predictions about how it applies to text-based or multi-party environments. Depending on how the mechanisms apply, they might also work for text processing and production. It does strongly predict that performance as well as what words are conventionalized should be highly dependent on matcher verbalizations (and not mere performance). It’s unclear how to extend it to a multi-person setup, it seems that alignment processes could take place is small group settings as well, but with only bottom-up processes, I think the identities of the conversational parties would get confused and lumped together, and who said what should not be a large factor. That said, Interactive Alignment allows for explicit social reasoning as a last resort when needed, so it mostly predicts substantially slower responses and more difficulty in circumstances where this is needed. It’s unclear how introducing a new listener affects the alignment; when alignment needs to start-over seems ill-defined, so I’m not sure what the predictions are for when an additional matcher is added, but the speaker is aligned with one matcher already.

RSA also hasn’t been fleshed out for these situations, but as it is build on social iterated reasoning, it should extend smoothly. In RSA, non-verbal feedback such as in time-to-selection or performance should be enough for speakers to update about the listeners understanding, and then converge to shorter utterances. Additionally, text-based communication should work similarly to spoken because they are mostly treated the same at the higher level of reasoning. As the number of participants and their levels of knowledge get more complicated, performance may degrade as speakers run into memory or computational limits and satisfice, but this should be a smooth gradual degradation and not an abrupt shift. Speakers should be able to take into account listeners differences and tailor utterances to the different knowledge levels. Depending on their goal (which may not be known), they may compromise between speed and being able to keep all listeners with them, so it may be strategic to leave a straggling or uncoorperative listener behind. The details of when this is true would be determined by the details of the utility function weights.

4 Multi-person interaction

Experiments on pragmatics have focused on communication is dyads; however, much communication takes place outside of dyadic contexts. For instance, we are taught in classrooms, socialize in groups, and work as teams. In these contexts, we formed shared knowledge as part of a group, with conventions or in-jokes that arise organically with the group, often from the contributions of more than two people. These interactions also often involve differing backgrounds, and so people navigate how to communicate to a group of people with varying unknown knowledge levels. In multi-person contexts, it is also possible to end up conveying information unevenly, creating conceptual pacts between some people while going over the heads of others. This is not desirable when the common goal is to have everyone understand a reference, as in typical reference games, but may be desirable in more realistic scenarios when not all goals are aligned.

This variability between participants is a key feature of mutli-person interaction; as listeners, each person may understand things differently, and with multiple speakers, listeners need to adapt to multiple people at once. These multi-person dynamics have been partially studied, with a focus on what happens when a director switches to talking to an individual matcher with a different set of knowledge.

4.1 Side-participants

Some iterated reference games have included more than two players; for the most part, they have focused on what happens when some people form a conversational pact and then another person joins in. In [Wilkes-Gibbs and Clark, 1992], a director and a matcher completed a sorting task, and then the director did that same sorting task with a different matcher. What varied was how much access the second matcher had to the initial pact formation; during the initial trials, the second matcher either sat next to the director, watched a video feed of the director from another room, overheard the director from across the room, or had not access to what was going on. The more access the matcher had, the more the director treated them as part of the pact and used short definite terms of reference, rather than long descriptions.

4.2 Mixed groups of matchers

A set of more recent work on how directors talk to multiple matchers comes from Sarah Brown-Schmidt's lab. All these experiments use a paradigm where the director and some matchers do a sorting task to establish a coordinated set of references. Then at test, the director does the picking task with some matchers, either ones who are knowledgeable, naive, or a mixed group.

In [Yoon and Brown-Schmidt, 2014], a director does the picking task with either just knowledgeable matcher 1 or with both a knowledgable matcher and a naive matcher. On each trial, 3 of the 4 image options are familiar tangrams and one is novel. When they are the only addressee, the experienced matcher looks at the novel tangram when the speaker is disfluent, presumably expecting hesistant utterances to go with the object without an established name. However, this expectation is overridden when there is a naive matcher present.

In [Yoon and Brown-Schmidt, 2019b],a director again does the picking task with either just knowledgeable matcher 1 or with both a knowledgable matcher 1 and a naive matcher 2. However, while the target object is the same tangram for both matchers, the context items differ. In some cases, the contexts are other tangrams from the sorting task, in other cases the target tangram, in other cases it's easily nameable clip art images. Directors take into account what information they need to convey for each matcher to get it right.

This provides evidence that directors track individual knowledge levels when there are two matchers, rather than averaging.

In [Yoon and Brown-Schmidt, 2019a], several experiments investigate the 'aim low' hypothesis. In experiment 1, the director forms individual pacts with two matchers in sequence. They then test with some combination of experience and naive matchers: either 1/0, 2/1, 1/2, or 0/1 (experienced/naive). As the proportion of naive matchers increase, the director's utterance length, rate of disfluencies, and rate of elaborations all increase. In experiment 2, training is again sequential, but they test the ratios 1/0, 3/1, 2/2, 1/3, and 0/1 (experienced/naive). The 1 experienced condition stands out, but all the conditions with at least half naive matchers pattern together.

By doing the training phases sequentially, even with the same director throughout, there's a risk that the director forms slightly different pacts with different participants. To control for this, in experiment 3, the director did the sorting task with three matchers at once. At test, the director talks with 3 matchers on the picking task (3/0, 2/1, 1/2, or 0/3 experienced/naive). There was a large difference between the all-knowledgeable condition and all the other conditions.

4.3 Worse than naive?

[Weber and Camerer, 2003] looked at performance with a worse-than-naive participant, one who had done the sorting task separately with a different partner. Dyads did the sorting task on office images, forming pacts picking out different key elements of each picture. At test time, a director did the sorting task with both their training partner and a person from a different original partnership. Utterances got longer after the new participant is added, and even knowledgeable matchers were slower. [Weber and Camerer, 2003] note anecdotal instances where the director and new matcher talked past each other about images because they each associated the image with a different key feature and were unwilling to accommodate the other's viewpoint.

5 Results

The major result of all of these papers is that in a variety of circumstances, people rapidly form conceptual pacts. Additionally, these facts can reform when the circumstances dictate, as the people or contexts of the conversation change. Thus, the pacts seem to be functional. This is consistent with an RSA style approach which would pressure behavior to be functional. It's unclear whether some of the mechanics are consistent with the Interactive Alignment account or not. However, this functionality is not absolute; the results of [Weber and Camerer, 2003] suggest that even when the pact shouldn't be in effect because of a new party; pacts may still serve as preconceptions that may get in the way of taking the perspectives of new conversational partners.

The results of studies with mixed-knowledge matchers are consistent with an RSA style model of social reasoning. Speakers produce different utterances when they are talking with only knowledgeable matchers versus when they need to help naive matchers. Matchers seem to expect this, and not be surprised when a new speaker or the addition of a naive matcher lead to descriptions inconsistent with the pact [Yoon and Brown-Schmidt, 2014, Metzing and Brennan, 2003]. Speakers track the different knowledge states of individuals, at least in small groups; giving short utterances if this will be clear to each listener in their context [Yoon and Brown-Schmidt, 2019b]. [Yoon and Brown-Schmidt, 2019a] shows that it's possible to form a conceptual pact with multiple matchers at once, and that it seems to work similarly to dyadic pacts; however, the dynamics of this process are not explored. These are all more consistent with theories of social reasoning such as RSA than with the Interactive Alignment account.

The other circumstance in which pacts can be broken is when the context changes. In [Ibarra and Tanenhaus, 2016], participants formed a pact about what they called the different plastic pieces they were identifying; in the next step, they used these pieces to construct Blokus animals. As the body part functions became clear, they switched to uses those terms instead. In a different experiment, names for objects were changed when new objects were uncovered that were also matches for the same name; rather than completely rename participants over distinguished the pieces with adjectives like ”real” or ”origami” to distinguish clip art animals from tangrams animals [Ibarra and Tanenhaus, 2016]. Overall these seem to indicate that parts are breakable when they cease to be functional, which would be consistent with an RSA view that the pact is being kept because it is is a contextually good choice, when it ceases to be because the context changes, there isn't a cost to breaking the pact. While the Interactive Alignment theory may explain how changes to the pact propagate, it does not explain why people can abruptly switch terms and break the pact.

6 Actual proposal

For my first year project, I propose to study the process of conversation pact formation within a triad of speakers. I will use the methods and analysis techniques of [Hawkins et al., 2020] to investigate how a director and two matchers coordinate on a sorting task. I will also experiment with whether changing roles within a group alters the conversation pact or speed of convergence. I propose to do one experiment where roles are constant and one person stays the director for the entire game, and a second experiment where the role of director switches halfway through the game, so one person is director/matcher, one is matcher/director, and the third remains a matcher throughout.

6.1 Methods

I will run this study using the cued matching set up from [Hawkins et al., 2020] where the director sees one of the tangrams highlighted and has to instruct the matchers to click on it. I'll use 12 tangram stimuli and 6 rounds. In the director-switch condition, the switch will occur after round 3; participants will not be told about the switch before it happens. Each participants will have a color symbol for the text chat, so they establish identities and can tell each other apart. Once each person has clicked, it will give feedback about who got it right to all three people (check or x next to color). I will use tangram stimuli both to increases comparability with [Hawkins et al., 2020], but also because the abstraction increases the chances that participants might end up in a position where they know that ”angel” refers to a certain tangram (based on an initial elaborated description) while thinking the tangram doesn't look like an angel. I'm curious whether if they become the director, they'll keep the pact and call it an angel, or go rogue and adjust to a term that makes more sense to them.

I will recruit participants online (from Mturk or Prolific) and host the games using the pipeline/framework described in [Hawkins, 2015].

6.2 Analysis Plan

I plan on comparing my data to that of [Hawkins et al., 2020] using the same tests. Additionally, I can take advantage of the additional matcher to compare matcher performance between and within triads, which may distinguish harder tangrams from bad descriptions. I'll be able to test the “aim low” theory be seeing what degree of utterance shortening occurs when one matcher is right and one is wrong compared to when both are right. I will also compare across the switch and no-switch experiments for changes in the accuracy and utterance length for the latter half of trials. In the switch experiments, I can also look at what changes to language occur at the director switch in terms of utterance length and content words.

6.3 Why this?

I want to do a fairly tight follow-up on [Hawkins et al., 2020] so I can use that as a scaffold to learn the techniques of running real-time reference games and analysing text data. I think that a conversation pact will successfully form in the interface because pact formation occurred between one director and several matchers in [Yoon and Brown-Schmidt, 2019a], but they don't analyse the dynamics of formation. A 2 versus 3 comparison between the results of [Hawkins et al., 2020] and a 3 person version of it would be a useful foundation to more work on multi-party language coordination.

Another seeming gap I'm trying to fill is between the set director-matcher roles that are usually used and the equal/collaborative roles used in different paradigms. By switching the director for another member of the pact, we can get an intermediate setting. In a 3-party we get a control on individual variability by looking at the performance of the 3rd person who remains a matcher. A number of more interesting multi-party coordination tasks will have multiple speakers with different knowledge, so checking how much a speaker sticks with the formed pact versus in-puts their own ideas will be a useful baseline for future work.

This experiment will generate a useful database of reference utterances which would be useful for modelling work.

7 Alternative ideas/Future work

I chose one piece of work that I think is an interesting and feasible starting point, but I'm curious about other variations that are beyond the scope of this project, but may be fodder for future work. I want to run a three-person version of [Hawkins et al., 2020]; I think it's also feasible to run some slight variation on that. As of October when I was writing my NSF, I was interested in whether pact formation still occurs when back-channel communication doesn't exist; I'm now more excited about switching out directors, but I'll include a brief sketch of no-feedback here, as a potential alternative extension.

One thing to test would be whether adaptation occurs when matchers cannot talk back and the only feedback is whether the correct image was selected. Comparisons between trials where matchers can communicate and where they cannot would clarify the role of matchers' utterances in converging on reference language. Notably the Interactive Alignment Account would predict that this would be challenging because there isn't any bottom-up alignment and speakers would have to use theory of mind reasoning. RSA models would predict less of a difference in performance based on modality of feedback. I expect that adaptation will occur even when matchers cannot communicate, but initial performance will be lower due to difficulty conveying initial reference. This pattern would support the RSA theory and also point to two-way communication being more important for establishing a communicative foundation than for refining terms. As an extension, I could test this further by showing new matchers the instructor utterances from a previous game to test whether understanding early trials is essential to understanding later shorthand references.

Also in the direction of altering levels of feedback, I wonder if you get pact formation if the speaker doesn't get feedback on matcher correctness, but the matcher does. This would encourage more matcher communication around what they understand and what they don't.

I'm also not proposing doing formal RSA-style modelling, although an eventual goal would be to adapt RSA-style computational models to multi-agent situations. A first step would be to take a simple computational model of teaching [Frank and Liu, 2018] and supplement it with NLP semantic models to map from words to meaning. I could then incorporate different features such as length, distinctiveness, and similarly to previous successful utterances, and test different feature weightings to see what fit human data.

Another random idea for stressing how much people model others separately would be to have a speaker alternately talk to multiple (siloed) matchers about the same set of tangrams; presumably they would try to form the same pact with each of them, but might end up with slightly different pacts and I wonder how well they could keep track.

One of the issues that seems to come up in theories is the need to explain why people prefer efficient utterances over long very specific ones. RSA deals with this by introducing cost terms, but it seems like there should be trade-offs between how long a speaker spends crafting the perfect utterance and how long or short the utterance is. While these games mostly don't introduce explicit time pressure; life comes with some implicit time pressure and participants have pressure to complete the task quickly so they can leave. There may be ways to push around the trade-offs of how pragmatic to be; for instance by limiting how many words a speaker can say, giving them a time window to think where they can't communicate yet, or in the opposite direction, by adding some form of time pressure or time limit.

References

[Baumann et al., 2014] Baumann, P., Clark, B., and Kaufmann, S. (2014). Overspecification and the Cost of Pragmatic Reasoning about Referring Expressions. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36.

[Brown-Schmidt et al., 2015] Brown-Schmidt, S., Yoon, S. O., and Ryskin, R. A. (2015). People as Contexts in Conversation. In *Psychology of Learning and Motivation*, volume 62, pages 59–99. Elsevier.

[Clark and Wilkes-Gibbs, 1986] Clark, H. H. and Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*.

[Frank and Goodman, 2012] Frank, M. C. and Goodman, N. D. (2012). Predicting Pragmatic Reasoning in Language Games. *Science*, 336(6084):998–998.

[Frank and Liu, 2018] Frank, M. C. and Liu, L. (2018). Modeling classroom teaching as optimal communication. Preprint, PsyArXiv.

[Garrod and Doherty, 1994] Garrod, S. and Doherty, A. (1994). Conversation, co-ordination and convention: An empirical investigation of how groups establish linguistic conventions. page 12.

[Garrod and Pickering, 2009] Garrod, S. and Pickering, M. J. (2009). Joint Action, Interactive Alignment, and Dialog. *Topics in Cognitive Science*, 1(2):292–304.

[Goodman and Frank, 2016] Goodman, N. D. and Frank, M. C. (2016). Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences*, 20(11):818–829.

[Hawkins et al., 2020] Hawkins, R. D., Frank, M. C., and Goodman, N. D. (2020). Characterizing the dynamics of learning in repeated reference games. *arXiv:1912.07199 [cs]*.

[Hawkins, 2015] Hawkins, R. X. D. (2015). Conducting real-time multiplayer experiments on the web. *Behav Res*, 47(4):966–976.

[Ibarra and Tanenhaus, 2016] Ibarra, A. and Tanenhaus, M. K. (2016). The Flexibility of Conceptual Pacts: Referring Expressions Dynamically Shift to Accommodate New Conceptualizations. *Front. Psychol.*, 7.

[Metzing and Brennan, 2003] Metzing, C. and 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.

[Pickering and Garrod, 2004] Pickering, M. J. and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behav. Brain Sci.*, 27(02).

[Weber and Camerer, 2003] Weber, R. A. and Camerer, C. F. (2003). Cultural Conflict and Merger Failure: An Experimental Approach. *Management Science*, 49(4):16.

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

[Yoon and Brown-Schmidt, 2014] Yoon, S. O. and Brown-Schmidt, S. (2014). Adjusting conceptual pacts in three-party conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(4):919–937.

[Yoon and Brown-Schmidt, 2019a] Yoon, S. O. and Brown-Schmidt, S. (2019a). Audience Design in Multiparty Conversation. *Cognitive Science*, 43(8):e12774.

[Yoon and Brown-Schmidt, 2019b] Yoon, S. O. and Brown-Schmidt, S. (2019b). Contextual Integration in Multiparty Audience Design. *Cognitive Science*, 43(12):e12807.