HOW EXTRA-LINGUISTIC GOALS AFFECT COMMON GROUND FORMATION

BY

KATE MARIE SANDERS

THESIS

Submitted in partial fulfillment of the requirements for the degree of Master of Arts in Psychology in the Graduate College of the University of Illinois at Urbana-Champaign, 2014

Urbana, Illinois

Adviser:

Professor Sarah Brown-Schmidt
The formation of common ground between interlocutors relies on their linguistic cooperation, manifesting as shortened references produced during a given conversation. However, people have goals for their conversations, and these extra-linguistic goals can influence a variety of linguistic choices. If extra-linguistic goals do affect the formation of common ground, then reference production length may vary with speaker cooperativeness. The present research addresses in three experiments: 1) the effect of extra-linguistic goals on reference productions, 2) the relationship between informativity of the reference productions and their length, and 3) how goal habituation affects references produced under different goal states. Consistent effects of condition on the reference production lengths are not observed, however, across Experiments 1a and 2, and the effects of goal habituation are inconsistent. Experiment 1b demonstrates that the length of reference productions is not related to informativity within the dialogue. However, consistent effects of round and instance are observed, as would be expected if the interlocutors were sharing conceptual pacts.
# TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION........................................................................................................1

CHAPTER 2: EXPERIMENT 1A........................................................................................................13

CHAPTER 3: EXPERIMENT 1B.....................................................................................................28

CHAPTER 4: EXPERIMENT 2.........................................................................................................36

CHAPTER 5: GENERAL DISCUSSION..........................................................................................47

CHAPTER 6: CONCLUSIONS........................................................................................................49

REFERENCES.................................................................................................................................52
**Chapter 1: Introduction**

Engaging in a conversation fills different purposes: for example, a person may need to argue a point successfully, persuade someone to give him something, or hide some knowledge from the other person. Having a conversation is a means to a desired end. That “end” may be completing a social nicety or getting some information from another person, perhaps, and simultaneously not giving away information you know. The linguistic choices made in the conversation can be tailored to meet such goals that are external to the conversation.

Various properties of speakers and addressees govern the descriptions that they will use in a conversation. For instance, the common ground, or mutually established knowledge, that the interlocutors share, and whether they have established a conceptual pact between themselves (Stalnaker, 1978; Clark & Wilkes-Gibbs, 1986; Brennan, 1996), affect the descriptors used. Similarly, the goals they share (Russell & Schober, 1999); their ability to perspective-take (Hanna, Tanenhaus, & Trueswell, 2003); and communicative pressures they are under, like time and complexity of the referents (Swets, Jacovina, & Gerrig, 2013), also impact the utterances that speakers design for their audience. We intend to investigate how forming common ground is affected by extra-linguistic conversational goals.

Theories of alignment – meaning developing shared ideas or representations – between interlocutors assume that interlocutors are being cooperative linguistically. Pickering and Garrod’s (2004) Interactive Alignment Account theorizes that production and comprehension are interwoven at multiple levels of communication, allowing for the automatic alignment of their word choice, grammar, prosody, and other linguistic representations. This theory relies on making forward models of one’s own production and on predicting the other interlocutor’s subsequent comprehension and production (Pickering & Garrod, 2013). This would mean that
the degree to which interlocutors share information – entrain on certain phrases – is thus determined by the covert and overt imitation of each other.

However, social linguistics studies show that linguistic choices can be influenced by extra-linguistic factors. For example, people employ audience design, changing their word usage, speech rate, and other qualities when the speaker believes that the audience expects such speech traits (Bell, 1984). Furthermore, having explicit task goals and separate beliefs about another person’s goals can affect the informativity and structure of a person’s conversation (Russell & Schober, 1999). Non-linguistic goals can affect the words chosen in a description, meaning linguistic choices can be influenced by external goals, such as social goals (Wardrow Lane & Liersch, 2011).

If extra-linguistic factors, such as goal states, do affect word choices, then entrainment may vary with speaker cooperativeness. We thus seek to answer the question: how does a task-based goal affect how interlocutors form common ground? Interlocutors may need to cooperate linguistically, but what if they are competing against each other in a task? The present research attempts to address this question by examining how the length and informativity of exchanges about shared items under discussion are affected by interlocutors’ extra-linguistic goals, in a task with both cooperative and competitive conditions in which pairs of interlocutors discuss shared items.

**Collaborative Effort in Communication**

Effective communication is built on shared references and common ground between interlocutors. Common ground is the shared knowledge between interlocutors, including common knowledge and items that are visually co-present, as well as references shared or developed between interlocutors (Clark & Wilkes-Gibbs, 1986; Pickering & Garrod, 2004). The
exact role that interlocutors’ shared knowledge plays in the production of a reference is heatedly debated (Horton & Keysar, 1996; Clark & Wilkes-Gibbs, 1986).

This assumption of cooperation is codified in Grice’s Cooperation Principle, which describes people’s linguistic goals (Grice, 1975). It holds that people cooperate in conversations so that the conversations are successful (Grice, 1975). To be understood, interlocutors must work together, and one way to do so is to form shared references to the objects under discussion (Clark & Wilkes-Gibbs, 1986; Garrod & Anderson, 1987; Brennan & Clark, 1996). Rather than just sending information (as the speaker) and receiving information (as the addressee), interlocutors collaborate to determine their shared terms (Clark & Wilkes-Gibbs, 1986). For example, sometimes speakers send insufficient information or need to correct what they have just said, and listeners may interrupt partway through to ask for clarification (Clark & Krych, 2004). This theory is a bilateral account of speaking and listening, in which they are a joint activity.

Pickering and Garrod (2004) develop a formalized version of the Cooperative Principle, including that people linguistically align with each other as they collaborate in a conversation, in their model of language processing reliant on low-level priming mechanisms. Their Interactive Alignment Account is a theory of how language production and comprehension could be interconnected, based on similarities to action and action perception. Pickering and Garrod (2013) argue that producers of language create forward models of their linguistic actions before producing words, and that comprehenders of language concurrently imitate the producer’s linguistic actions and subsequently model those actions. Such actions include predicting the semantic, syntactic, and phonological aspects of language production, much as the Interactive Alignment Account theorizes there are links at all levels of linguistic representation. Because of
the rapid covert imitation of the other person in a conversation, overt imitation – priming or lexical entrainment – is likely to occur.

One way of showing that interlocutors establish a shared acceptance of the terms they use is that referring expressions for the same discourse entity grow shorter in a conversation (Clark & Wilkes-Gibbs, 1986). Conversational partners will at first use many words to establish a specific reference, and then progressively need fewer words, because they rely on the prior shared acceptance of their terms (Clark & Wilkes-Gibbs, 1986). For example, two people may at first call an item under discussion “the flat spatula for flipping pancakes,” next call it “the flat pancake flipper,” and then continue calling it “the pancake flipper.” Entrainment results in consistently and repeatedly using the same phrase (or conceptualization) for an item under discussion (Brennan & Clark, 1996). Such entrainment over reference productions can be used as a measurement of their formation of common ground (Clark & Wilkes-Gibbs, 1986; Schober & Clark, 1989). Participants in a conversation try to minimize the effort needed to successfully produce and comprehend references; they follow the “principle of least collaborative effort” in conversation (Clark & Wilkes-Gibbs, 1986, p. 26). This principle states that speakers attempt to use phrases that are informative enough to accurately convey what they are referencing, without (inefficiently) using more information than necessary (Clark & Wilkes-Gibbs, 1986).

The “principle of least collaborative effort” explains why the same terms come to be used in a conversation: interlocutors minimize the effort needed for successful communication by continuing to use the same terms within a conversation. Garrod and Anderson (1987) found that entrainment increased – that is, two interlocutors came to agree more on terms used – during the course of completing a task together. Brennan and Clark (1996) found that when speakers talk to a new partner, they use variations on terms used in an immediately preceding discussion, rather
than commonly-used terms. This suggests that interlocutors work to establish terms together instead of just automatically using a simpler term.

Collaboration to form common ground between interlocutors creates a set of special knowledge between them. People not included in interlocutors’ collaboration do not understand the references produced as well as the collaborating interlocutors who produced them: the overhearers – non-participants – of a conversation were slower and less successful at the task than the active participants in the conversation were (Schober & Clark, 1989). Because interlocutors continually work to establish and maintain shared knowledge, according to the collaborative view, they have knowledge that overhearers do not (Schober & Clark, 1989). Furthermore, the more that a certain person is appears involved in a certain conversation, the more shared references the original participant in that conversation will use with the new person (Wilkes-Gibbs & Clark, 1992). Observers sitting next to a participant in the original conversation are treated more like an original participant – using the shorter references developed during the original conversation – than people who are listening in another room (Wilkes-Gibbs & Clark, 1992). Brennan and Clark (1996) argue that conceptual pacts are formed between interlocutors – that is, that they are agreeing how to “conceptualize” a term (p. 1484) – by showing that people use different references when talking about a particular object with one person than with another person not in the prior conversation. Thus, participants in a conversation have a special relationship: they share particular references in their common ground.

The length of reference productions within a dialogue is related to the formation of common ground and how informative a reference is. References are measured in duration and word count, which decrease with increased efficiency of reference production and increased common ground between interlocutors (Wilkes-Gibbs & Clark, 1992). Duration and word count
are typically used as a measure of the process of forming common ground (Wilkes-Gibbs & Clark, 1992). The Gricean concept of over- or under-informativeness can be evaluated based on the number of words used to describe an item. The first description of an item typically contains more words than later descriptions, because the speaker may use nonstandard noun phrases to initially establish understanding: the addressee requires more information to determine which is the correct item (Clark & Wilkes-Gibbs, 1986). If, for example, the addressee requires clarification of an item’s description, this indicates too little description – too few words – was provided, making that reference under-informative. Addressees assume that, for example, even prenominal adjectives convey some information, or else the speaker would not have included them (Sedivy, Tanenhaus, Chambers, & Carlson, 1999). Furthermore, if words are calibrated such that the amount of new information is roughly equalized across all the linguistic units in an utterance, there is still an implicit added value per word; that is, each word gives some information to the addressee (Jaeger, 2010). Content units are based on the words used, again assuming that each word adds some amount of information (Fussell & Krauss, 1992). If each word is information, then longer referential productions are more informative.

**Goals in a Conversation**

The influence of goals in conversation has been studied from linguistic, sociological, and social-psychological standpoints, and the resulting conclusions point towards a possible role for extra-linguistic goals in affecting the linguistic choices made in a conversation. This conflicts with the assumption that all conversations are basically cooperative endeavors. The types of goals – linguistic and extra-linguistic – and their possible effects are described further.

An interlocutor could simultaneously hold multiple goals for the outcome of a given conversation. Because conversations and communication in general do not take place in a
vacuum, the ever-present social and societal dimensions cannot be dismissed; an interlocutor can have individual and social goals, as well as the need to communicate effectively (Sarangi & Slembrouck, 1992). Interlocutors’ disparate aims could preclude their being helpful to each other in the conversation; thus, interlocutors’ cooperativeness should not be assumed (Sarangi & Slembrouck, 1992). For instance, a person may not want to stop and help another person who is lost in the city, and so his instructions to the lost person are not detailed enough to be effective. Furthermore, interlocutors can switch between types of relational goals in a conversation, such as pursuing an instrumental (task-related goal) or a self-interested goal (Keck & Samp, 2007). The progress of a conversation and the relationship between interlocutors can be affected by shifting goals. As such, calling interlocutors “cooperative” precludes the effects of social dimensions (Sarangi & Slembrouck, 1992). Cooperation at a linguistic level may not reflect social cooperation. This idea is very different from the assumption in psycholinguistic literature that conversationalists are cooperative (Clark & Wilkes-Gibbs, 1986; Garrod & Anderson, 1987).

Though linguistic cooperation can be closely tied to social cooperation, social cooperation is not necessary for linguistic cooperation (Lumsden, 2008). For instance, a person may want another person to believe an implicature that is not true; if person A asks person B, “Did you like the ornamental flamingo I sent you?” and person B replies, “My wife liked it as much as I did,” it is implied that both person B and his wife liked the flamingo. This is not necessarily the case. Such a conversation, in which the interlocutors have different goals – hearing that a gift was received warmly versus not revealing the true reception – demonstrates sharing linguistic cooperation without sharing the same goal. Both parties share the common ground of “the ornamental flamingo” in question. The non-linguistic aspects of a conversation – variously termed extra-linguistic, substantial, extra-communicative, and social – can be regarded
separately from the linguistic aspects of the conversation (Lumsden, 2008). This means that different goal states can be attributed to two interlocutors even though their conversation follows the typical Gricean maxims.

The extra-linguistic and linguistic elements of a conversation do, however, still interact. Interlocutors may alter the linguistic aspects of a conversation to fit their expectations of what the other interlocutors would understand more easily. People adjust their style of speaking to match another interlocutor (Bell, 1984). Speakers give more details to addressees based on speakers’ beliefs about how recognizable a person is (Fussell and Krauss, 1992). When shown pictures of well-known pop culture icons, people described them with fewer details than they described pictures of less-well-known people (Fussell & Krauss, 1992). Furthermore, speakers will make an effort to describe maps to addressees from the spatial perspective of the addressee, though less of an effort is made when that spatial perspective is rotated at odd angles from the addressee (Galati, et al., 2013). People take into account the expertise of others (Isaacs & Clark, 1987) and general beliefs about others when speaking with them. Linguistic choices in a conversation are affected by various concerns unrelated to the conversation itself.

Interlocutors also can take into account the goals of others, and non-linguistic goals can influence the linguistic choices in a conversation (Russell & Schober, 1999). For example, extra-linguistic goals can affect the amount of information shared by interlocutors. Russell and Schober (1999) explored how interlocutors balanced their beliefs about a partner’s goal with their understanding of the conversation by manipulating what information about the goal states the interlocutors had. Because speakers provided different levels of detail depending on what they thought the addressees’ goals were, Russell and Schober (1999) argue that the pairs designed their conversations with the stated or implicit goals in mind. The interlocutors’ beliefs
about the goal of their partners affected the language they used, specifically the type of
information they shared (Russell & Schober, 1999). Thus, the extra-linguistic aspects of a
conversation can impact the linguistic aspects of a conversation.

One type of goal an interlocutor may have is to not share privileged ground – information
only the speaker has (Grice, 1975) – with another interlocutor. Once privileged ground
information is shared, then that information may enter their common ground (Clark & Wilkes-
Gibbs, 1986). It may be the case that not all privileged ground information will become common
ground information even when shared (see Brown-Schmidt, 2012), but regardless, after sharing
information, the addressee may then know something previously known only by the speaker.
Sometimes, a person may not want to share information with another person. Wardrow Lane and
Liersch (2011) found that the amount of information revealed by speakers to addressees changes
depending on the incentive to keep certain information private. They found that a greater
incentive to keep information private led to a greater chance of sharing that information – that is,
using modifiers when they were unnecessary and implied what type of item was hidden
(Wardrow Lane & Liersch, 2011). This shows that speakers’ goals can indeed influence the
information they share (Wardrow Lane & Liersch 2011). In general, leaking information like this
is unlikely to harm a speaker’s communicative goal (Nadig & Sedivy, 2002); however, deception
is one instance when an extra-linguistic goal could be in conflict.

Ariga and Lleras (2011) suggest that goals are a form of cognitive control that are
susceptible to habituation. Over time, performance on tasks requiring a person to constantly pay
attention tends to worsen; if the task is a “vigilance” task, this is called a vigilance decrement
(Davies & Parasuraman, 1982). This change over time can be thought of as a failure of cognitive
control (Miller & Cohen, 2001; Ariga & Lleras, 2011). However, a person must sustain an
accurate goal representation in order to perform a task with a desired outcome, because a goal representation includes information about what needs to be done to meet that goal (Braver & Cohen, 2000). So, the goal must somehow be accessible and actively influencing one’s processing – but such constant exposure to the goal means that the cognitive control system suffers habituation effects (Ariga & Lleras, 2011). This means that the cognitive control system, like perceptual systems, comes to stop representing a stimulus after prolonged stimulation (Ariga & Lleras, 2011). Ariga and Lleras (2011) capitalize on the idea of briefly changing the stimulus in a vision task and thus avoiding habituation: when applied to the cognitive control system, momentarily deactivating the task goal does in fact avoid goal habituation and thus prevent a vigilance decrement. A consequence of such a vigilance decrement in conversation could be revealing a secret, as Wardrow Lane and Liersch (2011) found.

These arguments for an effect of goals on the level of information provided in a conversation suggest that extra-linguistic goals can affect linguistic choices. However, many theories about conversations assume that interlocutors are cooperative. Thus, we want to test whether linguistic choices change when speakers are being competitive rather than cooperative. We want to test in a straightforward way whether interlocutors with explicitly-stated goals use references to build common ground differently when those goals are different.

**Overview of the Experiments**

The goal of the present research is to determine the effects of extra-linguistic goals on reference production. If players’ goals do not affect their linguistic cooperativeness, then the process of mutually establishing references only depends on linguistic goals. So, there will be no significant differences in the formation of common ground (as measured by exchange duration and word count) between competitive pairs and cooperative pairs. This outcome follows the
Cooperation Principle and the mechanistic approach to language processing and reference production theorized by Pickering and Garrod (2004), because their model of language processing relies on low-level priming mechanisms, and thus predicts that interlocutors’ extra-linguistic goals will not affect their exchanges.

Alternatively, if players’ extra-linguistic goals do affect their reference productions, then some difference in their exchanges will be found between conditions. One possible difference is suggested by the assumption that competitive people are working to be unhelpful to each other: they would use less-informative exchanges to refer to tangrams. This hypothesis predicts that the competitive exchanges will be shorter, because they are less informative. Competitive players would be more likely to produce references that are unclear or not useful to the other player; they might provide less information because they think that the other person is also providing less information in an attempt to make the game harder for them to win. Conversely, the cooperative exchanges will be longer, erring towards more rather than less information. Being over-informative is not ideal, but it is easier for a listener to recover from than from an under-informative description; over-informative descriptions are not always harmful to a listener’s understanding nor always rated as worse descriptions (Engelhardt, Bailey, & Ferreira, 2006). So, players attempting to be helpful – cooperative players – would give too much information, rather than too little.

The present research will address these hypotheses. Experiment 1a will test whether there are differences in reference productions between competitive interlocutors and cooperative interlocutors, using a task with either a competitive or a cooperative goal in which the interlocutors will refer repeatedly to the same objects. The goal of the experiment is to determine whether a difference between goal states affects reference production. Their referring
expressions will be analyzed for the number of words used to describe the items under discussion and the duration of their exchanges.

Experiment 1b will test the informativity of the exchanges produced, by presenting a subset of the exchanges to non-participants of the game. These listeners – functioning as “overhearers” of the exchanges, as they could not ask questions of the original speaker (Schober & Clark, 1989) – rated which item they thought was being discussed. The accuracy of the ratings is a measure of how informative the exchanges were. This will test the informativity hypothesis.

Finally, Experiment 2 will attempt to replicate the findings of Experiment 1a and demonstrate that reference production is affected by extra-linguistic goals, utilizing the theory of goal habituation. If people habituate to extra-linguistic goals, then the patterns of reference production will be affected by breaks during goal maintenance (Ariga & Lleras, 2011). Specifically, the pairs’ exchanges will revert to the start-of-game lengths after breaks in the game, because the goals will be re-sensitized; pairs’ performances will resemble their initial levels, when their goals were first established. Experiment 2 will thus test in another way whether goals affect how interlocutors come to refer to items under discussion.
Chapter 2: Experiment 1a

In this experiment, the task is a game that requires participants to refer repeatedly to the same objects, so that their exchanges about these objects could be tracked over time. During the game, participants take turns asking and answering questions about the locations of objects. The conditions have identical gameplay, but the outcome of the game is framed differently at the beginning and reinforced throughout the game. The identical gameplay is crucial, because any differences in the lengths of the exchanges produced between the conditions could not be attributed to the players’ doing different tasks.

Method

Participants

Thirty-two undergraduates at the University of Illinois at Urbana-Champaign participated for course credit. All were native English speakers and naïve to the experimental manipulation.

Materials & Design

The participants (hereafter, Player 1 and Player 2) played the game with physical pieces at separate desks, facing each other with a barrier set between them such that the players could hear but not see each other. This setup facilitated recording each session with a video camera, for use in later analyses (see Figure 1a for the layout). Players wore headsets with microphones that recorded their conversation onto a DAT solid-state recorder.
Figure 1a. This diagram illustrates the layout of the experiment. The two players face each other separated by a divider, with the experimenter in the room (at the computer). Their boards are identical (except for ship locations).

Figure 1b. An example board with tangrams affixed. The blue ships adjacent to certain tangrams were the player’s ships.

The game was played on two separate game boards, one for each player. Each game board was an 11x14” white dry erase board with five tangrams printed on cardstock and taped on the board at pre-arranged locations (see Figure 1b). The tangrams were abstract shapes resembling humanoid figures able to be described with varying difficulty. In each round, the players received the same exact boards. Between rounds, each pair of boards was switched for a new pair of identical boards.

Four blue magnets with pictures of blue ships attached were placed at pseudo-randomly assigned locations around the tangrams (such that no more than one ship was at any one location, and at least one player had at least one ship at each tangram). There were 20 total possible ship locations (one per each of the four sides of the five tangrams). Player 1’s ship locations were pseudo-randomly selected with no reference to Player 2’s ship locations. Each player had a separate set of magnets, including four red magnets with red ships glued on them and 18 magnets
of a different color (purple or green). The players used these to mark locations on the game board during the game.

Five pairs of boards in total were constructed, each pair with five different tangrams, for a total of 25 different pictures. The tangrams differed between the five boards, but the same five locations on the board were always used (see Figure 1b). Each pair of boards had identical tangrams, but different ship placements. One board from each pair was always used by Player 1, and the other board of the pair was used by Player 2. The order in which the five boards were presented varied according to a counterbalanced list; there were four presentation orders, including two that were determined randomly and two that were the reverse of those orders. Both conditions received the presentation orders equally. Each pair of participants was randomly assigned a condition. The conditions were compared between participant pairs, with eight pairs in each condition.

Procedure

The participants entered the room and sat at the desks facing each other, unable to see each other. They were randomly assigned their roles (Player 1 and Player 2). Both players alternated within each round as the director asking about a tangram and the matcher answering.

The experimenter explained the game to the players: they were participating in a communication task together, playing a game similar to the common game of Battleship, in which each player had a game board with ships on it. Players in both conditions were informed that a round ended when all four of one player’s ships were found. The players were told that the blue ships had been placed randomly on each board, and to find them, the players would take turns asking whether the other player had a ship at a certain location.
Players were shown where the ships could be located on a board. The ships were situated near tangrams, in any of four possible locations: above, below, to the right, and to the left of each tangram. The players had to ask about the ships’ locations relative to a tangram; for example: *is there a ship to the right of the picture of the woman running?* They were also informed that one of their ships could be at the same location as the other player’s in a given round.

The next set of instructions involved how to actually play the game. The players were instructed to follow a question-answer format, but told that they could ask the other player for clarification if needed. They could not ask for locations based on the placement of the tangram on the board, such as *the picture on the right side or the middle picture*; players were instructed to use the tangrams and not their placement on the boards. Players were instructed to mark ship locations with their magnets as they guessed: the red ships were used for marking “hits” where a player found a ship, and the other magnets were used for marking “misses” where the player guessed a location where there was no ship.

Example dialogue with an accompanying demonstration was provided to explain the steps of the game: Player 1 (acting as director) asks a question, Player 2 (matcher) answers, Player 1 marks his or her board, and then the process repeats, but with Player 2 next being the director. An example of asking for clarification during gameplay was included. Players were encouraged to ask the experimenter any questions they had about gameplay at any time.

Importantly, the instructions for the different conditions put this gameplay in two different contexts. “Competitive” pairs were instructed: “You are *competing* with each other to find ship magnets. You want to find the other person’s four ships *before they find yours*.” In contrast, “Cooperative” pairs were instructed: “You are *cooperating* with each other to find ship magnets. You want to find the other person’s four ships or you want them to find all four of
These tasks had the same outcome – all four of a person’s ships needed to be found in order to end the game – but the different conditions elicited different goals. Players were reminded of their goal at the end of the instructions.

Another crucial difference between conditions came at the end of each round. After each round in the competitive condition, the player who found the four ships was congratulated. After each round in the cooperative condition, both players were congratulated for finding the ships. The players’ goal state was thus reinforced throughout the game.

The game began after the digital audio recorder and the video camera were turned on and the experimenter told the player assigned to start the game. Gameplay proceeded as described in the instructions to participants. When all four of a player’s ships were found, the experimenter congratulated the appropriate person(s) and switched out the boards for the next set, following the assigned presentation order.

Occasionally during the games, players made mistakes, some of which necessitated the experimenter’s intervention. If players asked about tangrams using the tangram’s location on the board or named tangrams with a variant of the same one you just asked about, then the experimenter reminded them at the end of the round that such names were not allowed. Also, the players occasionally made mistakes in which one player’s information did not match what actually was on the board. For instance, Player 1 asked about a certain location, Player 2 said a ship was there, and Player 1 marked that location as having a ship, despite there actually being no ship there on Player 2’s board. During gameplay, if the players noticed a mistake, they generally fixed it through further clarifications. If they asked the experimenter, she told them attempt could try to fix their miscommunication.
After the players completed the five rounds, the overall winner was congratulated (for the competitive condition) or both players were congratulated for finishing (cooperative condition). Both players were debriefed about the purpose of the experiment and the different conditions.

Predictions

For this experiment, there are two potential outcomes. If players’ goals do not affect their linguistic choices, then there will be no significant differences in the exchange durations and word counts between competitive pairs and cooperative pairs. The second outcome is that there is some difference in competitive and cooperative exchanges, because players’ extra-linguistic goals do affect their reference productions. If so, it is likely that the competitive exchanges will be shorter, as the competitive interlocutors are working to be unhelpful to each other by using less-informative exchanges.

Data Analysis

Four undergraduate assistants, naïve to the experimental conditions, transcribed the participant recordings. Interrater reliability for the words transcribed from the players’ entire exchanges was measured with Fleiss’s kappa, \( k = .869 \). The assistants transcribed both the individual tangram references and the players’ entire exchanges used to establish reference to a given tangram. They labeled players’ references to tangrams, to provide a measure of noun phrase duration and a word count. A reference included any words used to describe a tangram. Listed below are two examples of exchanges that were included in the analyses:

Ex.1.

P1: the person who looks like they’re running with the two triangle feet closer together
P2: closer together and leaning towards the right?

Ex.2.

P2: the picture that looks like a long base to the left and three triangle types to the right
The assistants transcribed the words of each exchange, excluding filler words. These transcriptions were used to obtain the durations of the initial reference to the tangram within the exchange. Because the length of the entire exchange was of interest, the author subsequently marked the entire exchanges about each tangram, with their word count including all words coded to describe a tangram, and with duration including all pauses during reference production. These references were used for the current analyses. The assistants also transcribed the entire recordings separately, but these are not being used in the current analyses. The exchanges coded were analyzed for word count and duration (as done in Wilkes-Gibbs & Clark, 1992).

By examining the video recording, the author matched these exchanges to the actual tangrams named by players. The exchange heard was attributed to the tangram around which the director placed a magnet. For instance, in Ex. 2, the director asked about the picture that looks like a long base to the left and three triangle types to the right and placed a magnet by a particular tangram after hearing the matcher’s answer, so that exchange was coded as referencing that particular tangram. Mistakes were also coded: if a magnet placed on the board did not match the information on the other player’s board, then a mistake was recorded with that tangram name. For example, if a position held a blue ship on the director’s board, and the matcher marked that position on his or her board with a “missed guess” magnet, then that communication was considered a mistake. Occasionally, the players did not reference a tangram with any name, using phrasing like do you have a ship there? as their question. These could not be counted as exchanges, because they were not noun phrases describing tangrams; there were 26 of these exchanges that were not analyzed. Players were discouraged between rounds from using such phrases.
The author further coded (blind to condition) the tangram references into different types of reference productions. Names used outside of the question-answer format, such as asking the experimenter a question or attempting to fix an error, were both excluded from further analyses; there were a total of 37 non-gameplay exchanges. They were excluded because these names were not part of the “conversation” between the participants and thus did not reflect gameplay.

Many tangrams were referenced more than once throughout a round. To evaluate how the referential process changed with repeated reference, we coded exchanges for which number “instance” the reference was. This was used to determine the change over time in the repeated references to the same tangram. The number of instances decreased because not all tangrams were referred to an equal number of times during the task; that is, nearly all tangrams were referenced at least once by each pair, but few (73 instances) were referenced greater than seven times. This is due to the nature of the game: it is likely that all four of a person’s ships are found before the players exhaust all the locations around all the tangrams. All exchanges produced past the seventh instance were thus binned with the seventh instance, so that the numbers of instances per level were roughly equal.

**Results**

To test whether the players’ goal state affected their reference production, we measured the lengths of their exchanges (both duration and word count) across each round of the game and compared them between conditions. Mixed effects models were used for the comparisons.

**Accuracy**

Before comparing the exchanges, however, we first checked that the players had performed well at the task. The mistakes made between players in each condition were examined to determine whether players had both sufficiently understood the task. Mistakes were measured
as occurring when the director placed their ship magnet incorrectly. Players misunderstood which tangram was being discussed in 2% of all exchanges (43 out of 2170); these mistakes were identified as when the director put a Hit magnet where the matcher did not have a ship or when the director did not put a Hit magnet where the matcher did have a ship, indicating that the players were discussing different tangrams. Players made mistakes unrelated to communication errors, specifically, forgetting to mark the location they had just asked about, but such mistakes occurred in 2% of exchanges (33 out of 2170); these mistakes were identified by watching the gameplay from the video recording. So few mistakes indicated that the players were competent at the task. These “mistake” trials are included in further analyses of the exchanges.

*Duration*

The average duration of the exchanges across rounds is depicted in Figure 2. Inspection of Figure 2 showed differences in the overall performance of the conditions across rounds: the average exchange durations decreased for cooperative pairs (4.50 to 3.91 sec) and increased for competitive pairs (2.88 to 4.20 sec), and the differences between conditions generally diminished across the rounds; both slopes differ from zero. The exchange durations were explored with a mixed effects model described in Table 1.
Figure 2. Average duration over rounds by condition. Cooperative players produced longer exchanges on average in the first two rounds than in the final three, compared to competitive players.

To build the mixed effects model used to predict duration, I used the maximal random effects structure (Barr, Levy, Sheepers, & Tily, 2013). The model included the random intercepts by subject, pair, and tangram, and random slopes for round and condition as appropriate. Condition (competitive or cooperative) and Round (1-5) were included as predictors. Both factor contrasts were mean centered. The full model included the random effects of pair, subject, and tangram, each as random intercepts and all possible random slopes. The full model failed to converge. The best-fitting model was found by using a backwards-fitting procedure; it included all random slopes except for tangram as a random intercept (Barr, et al., 2013). Instance was coded with seven levels; all instances above the seventh instance were binned with the seventh, as described. The p-values were calculated for the fixed effects using the deviance test statistic from model comparison (Barr, et al., 2013).

The durations within rounds were compared between conditions. Visual inspection of Figure 2 showed that the average duration of the first instance – the first exchange in which a
particular tangram was discussed – was longer for the cooperative condition (10.40 sec) than the competitive (7.70 sec). The model output is displayed in Table 1. There was a main effect of instance \((t=-11.89, p<.001)\); this means that overall, the participants shortened their terms with each referral. This interaction between condition and instance was marginal \((t=-2.19, p=.10)\). This comparison shows that cooperative players produce longer references than competitive players when each tangram is first discussed. The interaction between condition and round was significant, \(t=2.21\) \((p<.05)\). The condition effect at each round was tested and was significant at Round 1 \((t=-1.73, p<.05)\) and Round 3 \((t=-1.68, p<.05)\); it was marginal for Rounds 2 and 4. The condition effect was thus not present at the later rounds of the game.

A supplemental analysis was conducted to determine whether the initial reference, rather than the full exchange – including back-and-forth between the players – drove these effects. The model was tested using the initial references’ durations as the dependent variable, rather than the exchanges’ durations. The interaction between condition and round that was found for exchange durations was not present for just the initial references’ durations, \(t=-1.53\) \((p>.05)\). This suggests that the initial references did not solely drive the differences in exchange durations across rounds, but rather the conditions differed on the length of the full exchanges.
Figure 3. Average duration of tangram exchanges across instances, using all rounds and tangrams. The 9th instance of a tangram being referred to occurred thrice in the cooperative condition when players inadvertently asked about locations they had already marked with ships.

Table 1
Parameter estimates by subjects, tangrams, and pairs for analysis of condition, round, and instance effects in duration. Significance tests for fixed effects were estimated using model comparison. The data analyzed included 2170 exchanges from 32 subjects (16 pairs) and 25 tangrams.

|                          | Estimate | SE  | z-value | Pr(>|z|) |
|--------------------------|----------|-----|---------|----------|
| Fixed effects            |          |     |         |          |
| (Intercept)              | 3.19     | 0.21| 14.93   | .001     |
| Condition                | -0.27    | 0.19| -1.37   | .166     |
| Round                    | -0.04    | 0.11| -0.33   | 1        |
| Instance                 | -1.09    | 0.09| -11.89  | <.001    |
| Condition*Round          | 0.23     | 0.11| 2.21    | .030     |
| Condition*Instance       | 0.20     | 0.09| 2.19    | .100     |
| Round*Instance           | -0.11    | 0.06| -1.91   | .168     |
| Condition*Round*Instance | -0.05    | 0.06| -0.91   | 1        |

Random effects

Groups

<table>
<thead>
<tr>
<th></th>
<th>Variance</th>
<th>Std Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Intercept)</td>
<td>0.18</td>
<td>0.43</td>
</tr>
<tr>
<td>Instance</td>
<td>0.05</td>
<td>0.22</td>
</tr>
<tr>
<td>Round</td>
<td>0.09</td>
<td>0.30</td>
</tr>
<tr>
<td>Instance*Rd</td>
<td>0.04</td>
<td>0.21</td>
</tr>
<tr>
<td>Tangram</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Intercept)</td>
<td>0.20</td>
<td>0.45</td>
</tr>
<tr>
<td>Pair</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Intercept)</td>
<td>0.32</td>
<td>0.56</td>
</tr>
<tr>
<td>Instance</td>
<td>0.05</td>
<td>0.24</td>
</tr>
<tr>
<td>Round</td>
<td>0.03</td>
<td>0.18</td>
</tr>
<tr>
<td>Instance*Rd</td>
<td>0.004</td>
<td>0.18</td>
</tr>
</tbody>
</table>

*p<.05
**Word Counts**

The word counts within rounds were likewise compared between conditions using a Poisson distribution, binning Instance as before. Figure 4 showed that the average word count of the first instance was longer for the cooperative condition (21.86 words) than the competitive (17.96 words), similar to the pattern of results for duration. The model output is displayed in Table 2. The interaction between condition and round was marginal, $z=1.91$ ($p=.057$). The best-fit model indicated there is a main effect of instance ($z=-10.87$, $p<.001$), indicating that the speaker and addressees’ exchanges shortened with repeated reference. The interaction between condition and instance was significant ($z=2.16$, $p<.05$), though it was only marginal for duration. The effect of condition on instance was explored, and there was a marginal effect of condition on the first instance ($z=-1.73$, $p=.08$), but no significant effects. This indicates that variation between conditions is due to the first instance. Unlike the analysis of duration, the interaction between round and instance was significant for word count ($z=-3.67$, $p<.001$). The rounds were compared by instance; the effect of round was significant for the first instance ($z=3.14$, $p<.01$), but was not significant at the other instances, showing that the first instance drives the round by instance interaction. The instance of a tangram exchange thus accounts for the most variance between conditions across rounds.
Figure 4. Average duration of tangram references across instances, using all rounds and tangrams. The 9th instance of a tangram being referred to occurred thrice in the cooperative condition when players inadvertently asked about locations they had already marked with ships.

Table 2
Parameter estimates by subjects, tangrams, and pairs for analysis of condition, round, and instance effects in word count. The data analyzed included 2170 exchanges from 32 subjects (16 pairs) and 25 tangrams.

| Fixed effects          | Estimate | SE  | z-value | Pr(>|z|) |
|------------------------|----------|-----|---------|---------|
| (Intercept)            | 2.01     | 0.07| 29.89   | .309    |
| Condition              | 0.06     | 0.06| 1.01    | .293    |
| Round                  | -0.03    | 0.03| -1.05   | .293    |
| Instance               | -0.25    | 0.02| -10.87  | .057    |
| Condition*Round        | 0.06     | 0.03| 1.91    | .031*   |
| Condition*Instance     | 0.05     | 0.02| 2.16    | .001*   |
| Round*Instance         | -0.03    | 0.01| -3.67   | .281    |
| Condition*Round*Instance | 0.01    | 0.01| 1.08    | .281    |

<table>
<thead>
<tr>
<th>Random effects</th>
<th>Variance</th>
<th>Std Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subject</td>
<td>(Intercept)</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>Instance</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>Round</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>Instance*Rd</td>
<td>0.001</td>
</tr>
<tr>
<td>Tangram</td>
<td>(Intercept)</td>
<td>0.020</td>
</tr>
<tr>
<td></td>
<td>(Intercept)</td>
<td>0.056</td>
</tr>
<tr>
<td></td>
<td>Instance</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>Round</td>
<td>0.012</td>
</tr>
<tr>
<td></td>
<td>Instance*Rd</td>
<td>0.0005</td>
</tr>
</tbody>
</table>

*p<.05
One possible contributor to the cooperative pairs’ longer exchanges is that the addressee asked for clarification (or otherwise interacted with the speaker) in significantly more exchanges in the cooperative condition than in the competitive condition, \( z = 2.61 \) (\( p<.01 \)). However, there were 104 cooperative clarifications and 62 competitive clarifications, together accounting for only 8% of all exchanges.

**Discussion**

Experiment 1a tested whether or not the length of the exchanges was affected by the interlocutors’ goal. Overall, participants in both conditions shortened their exchanges over repeated references to the same tangram, like in Brennan & Clark (1996). The pattern of average exchange lengths differed between conditions at the beginning of the game, then converging towards the end. The average exchange lengths also differed at the first exchange about a new tangram: competitive players had shorter durations and word counts than cooperative players. It appeared that the instance was the greatest source of variance, particularly the first instance, which affected both exchange durations and word counts.

This pattern follows the predictions of the informativity hypothesis, because the competitive players use fewer words and less time in their exchanges about the tangrams, as if being less helpful or under informative. However, it is not directly known from this data how much information was contained in each exchange – perhaps the competitive players were more specific in their descriptions, though this contradicts the idea that competitive players put less effort into establishing shared references with each other. Experiment 1b addresses the question of informativity.
Chapter 3: Experiment 1b

The goal of Experiment 1b was to test how much information about the tangrams was provided by the Experiment 1a interlocutors in their exchanges. Specifically, we examined how accurately “overhearers” of the exchanges selected the tangram that the original speaker(s) described. This tests the informativity hypothesis by using the accuracy with which a person who was not one of the establishers of the shared reference can select the tangram discussed. A more informative exchange should be more understandable to an overhearder than a less informative exchange, and thus its referent guessed with greater accuracy. If the longer exchanges between the speaker and listener were also the more informative exchanges, we would expect that the greater amount of information in the longer exchanges would allow overhearers to more accurately select the described tangram after hearing a longer exchange.

The full exchanges were used rather than just the director’s initial reference to any given tangram, because Experiment 1a found that the referring differences between conditions were found at the level of the exchange, not just the initial reference within a given exchange. The first exchange about a given tangram – the first instance – was used because the greatest difference in durations between conditions was found for the first instance.

Method

This experiment was conducted over Mechanical Turk, or “MTurk,” which is an online system that connects potential workers (“Turkers”) with tasks, called “HITs,” that are provided by “requesters.” Turkers select tasks from the online marketplace and complete them for a small fee; the acceptable rate for compensation is about half-a-cent per question (Schnoebelen & Kuperman, 2010). Previous studies (Schnoebelen & Kuperman, 2010; Sprouse, 2011) have used
MTurk to collect similar linguistic data from Turkers, and Turkers’ performance closely replicates studies utilizing typical subject pools (Crump, McDonnell, & Gureckis, 2013).

Participants

A total of 203 workers from Mechanical Turk participated in the experiment; 119 females and 84 males participated. Participation was restricted to those from the USA with an approval rate of at least 95% (requesters on Mechanical Turk approved 95% or more of that participant’s previous submissions). Only native speakers of English were tested; if participants did not rate themselves as native English speakers, their data were discarded. Participants received $1.10 and were informed that the study would take approximately 45 minutes; participants took about 40 minutes to complete their HIT. Surveys were considered incomplete if more than five questions were left unanswered, and incomplete surveys were discarded prior to any analysis.

Materials

A total of 398 exchanges from Experiment 1a were presented to the Mechanical Turk participants (hereafter called “Turkers”). Each exchange was about the first instance for each of the tangrams; that is, it was about the first time that specific tangram was mentioned in a round. All first-instance exchanges were presented. The 16 pairs of participants in Experiment 1a encountered 25 different tangrams. There were thus 400 possible first instances, but one cooperative pair did not talk about two tangrams, resulting in the 398 first-instance exchanges used. The exchanges used consisted of the description of the tangram under discussion. Most exchanges (284) only included the speaker, but some (114) included the addressee’s request for clarification and the speaker’s answers. These exchanges were the original audio files, and so the speaker’s (and addressee’s) disfluencies and pauses were included.
The exchanges were divided between three surveys with 100 questions each and one with only 98 questions. Each Turker therefore listened to about 100 audio files and responded. The 398 exchanges were divided into the four surveys because it seemed unlikely that participants would respond to a survey to the best of their ability if it lasted considerably longer than an hour.

How the 398 exchanges were split between the four surveys was pseudo-randomly determined, such that half of the exchanges in each survey were cooperatively-produced, and half were competitively-produced. Turkers were not told which was which. The four resulting surveys were always presented to Turkers in the same order. Each exchange was paired with the five tangrams that were present when the exchange was originally produced. The order of the tangrams on the screen was randomized.

In the surveys, participants were instructed to click the audio file to hear the description and then “pick the image (1-5) you think is being described.” Each survey question contained an audio file and five tangram pictures, labeled 1-5, followed by the question, “which image is being described?” Participants then listened to the audio file and selected their answer. Participants could move forward and backward freely in the survey.

At the end of each survey, the participants were asked some demographic questions, including their gender, nationality, age, overall ability in the English language, and number of fluently-spoken languages.

Procedure

A link for a web-based survey through SurveyMonkey, an online questionnaire survey company, was posted in an MTurk HIT. Turkers were required to give their informed consent in order to take the survey. Turkers also had to pass an audio file check, to determine whether they
could adequately hear the audio files presented to them. At the end of the survey, participants were given a unique code to enter into the MTurk HIT to verify that they completed the study.

Predictions

The accuracy of the Turkers’ selected tangrams will be used to test the informativity of the exchanges. Informativity of an exchange is hypothesized to relate to the length of the exchange – using more words is equivalent to being more informative. Following this, if longer exchanges are more informative, then Turkers’ accuracy will be better following longer exchanges. This is because Turkers have no other information about the tangram set than the exchange presented, so their accuracy depends only on the information within the exchange. If the competitive players are being under-informative, then Turkers’ accuracy will be worse for competitively-produced exchanges than for cooperatively-produced exchanges. This is because Turkers will have less information from exchanges produced by competitive players than from cooperative players.

Results

The dependent measure used in our analyses is informativity, which is operationally defined as the Turkers’ accuracy at identifying the tangram to which the original speaker was referring. Preliminary analysis showed that Turkers’ average accuracies were not different between conditions: they were 84% correct on competitively-produced exchanges and 83% correct on cooperatively-produced exchanges. (By contrast, the original speakers and addressees mutually understood 92% of these same exchanges.) By round, Turkers were more accurate for exchanges produced in rounds on, three, and five (average of 86%, 85%, and 87% accurate, respectively) than for rounds two and four (average of 81% and 80% accurate, respectively). The Turkers’ average accuracy was then compared between conditions across the rounds in which the
exchanges were originally produced, because the Experiment 1a results showed differences in performances between conditions across rounds. The average accuracies across rounds (shown in Figure 5) follow different patterns between conditions.

Informativity was analyzed with a maximal mixed effects model fit with the logit-link function and accuracy as the dependent measure (right or wrong). The fixed effects were rounds, coded with orthogonal contrast codes; and condition, centered between levels. For this and subsequent analyses, when the maximal model did not converge, the random slopes were removed one by one until convergence was reached. The best-fitting model included participants and tangrams as random slopes by round and condition; it is shown in Table 3. There was a significant round by condition interaction ($z = 3.10, p<.05$).

**Table 3**
Parameter estimates by subject and tangram for analysis of condition and instance effects on accuracy. The data analyzed included 20196 responses from 203 subjects.

| Parameter              | Estimate | SE    | z-value | Pr(>|z|) |
|------------------------|----------|-------|---------|----------|
| Fixed effects          |          |       |         |          |
| (Intercept)            | 2.01     | 0.14  | 14.86   | <.001*   |
| Condition              | -0.22    | 0.16  | -1.33   | .182     |
| Round                  | -0.03    | 0.04  | -0.85   | .393     |
| Condition*Round        | 0.29     | 0.09  | 3.10    | .002*    |

<table>
<thead>
<tr>
<th>Random effects</th>
<th>Variance</th>
<th>Std. Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subject (Intercept)</td>
<td>0.675</td>
<td>0.82</td>
</tr>
<tr>
<td>Round</td>
<td>0.0001</td>
<td>0.01</td>
</tr>
<tr>
<td>Condition</td>
<td>0.003</td>
<td>0.06</td>
</tr>
<tr>
<td>Cdn*Rd</td>
<td>0.008</td>
<td>0.09</td>
</tr>
<tr>
<td>Tangram (Intercept)</td>
<td>0.357</td>
<td>0.60</td>
</tr>
<tr>
<td>Round</td>
<td>0.032</td>
<td>0.18</td>
</tr>
<tr>
<td>Condition</td>
<td>0.611</td>
<td>0.78</td>
</tr>
<tr>
<td>Cdn*Rd</td>
<td>0.190</td>
<td>0.44</td>
</tr>
</tbody>
</table>

*p<.05

Exploring the interaction revealed that accuracy for the competitively-produced exchanges about tangrams significantly decreased across rounds ($z = -2.89, p<.05$), but accuracy
for the cooperatively-produced exchanges marginally increased ($z = 1.85, p = .065$). The interaction is driven by the decrease in accuracy for competitively-produced exchanges from the first round to the second round, as the condition effect is significant at rounds one ($z = -3.23, p < .01$) and two ($z = -2.71, p < .01$) but is not significant at the other rounds.

![Figure 5: Comparison between accuracy at selecting tangram discussed in competitively-produced or cooperatively-produced exchanges, shown across the rounds in which the exchanges were originally produced.](image)

The duration and word count of the original exchanges were tested separately as predictors for accuracy of selecting the tangram being referred to, but neither were significant predictors. When included as a fixed effect with round, neither duration ($z = .46, p = .644$) nor word count ($z = -.46, p = .643$) was a significant predictor; when included as a fixed effect with condition, neither duration ($z = .02, p = .987$) nor word count ($z = -.05, p = .293$) was a significant predictor of accuracy. Overall Turker accuracy (the proportion of Turkers who correctly identified the tangram) and duration were not correlated ($-.07$), and word count and Turker accuracy were also not correlated ($-.06$). Splitting the data between competitively-produced
exchanges and cooperatively-produced exchanges did not reveal any effects of condition; for competitively-produced exchanges, accuracy was not correlated with duration (-.08) nor with word count (-.08), and for cooperatively-produced exchanges, accuracy was also not correlated with duration (-.05) nor with word count (-.04). Given that the respondents’ accuracy is a measure of how informative the exchange is, these findings indicate that exchange length is unrelated to the informativity of the exchange.

Discussion

Experiment 1b tested whether or not the conditions’ performance was linked with the informativity of the exchanges. According to the informativity hypothesis, the longer exchanges should be more informative; however, Experiment 1b showed no effect of exchange length on accuracy. The hypothesis was not supported. The accuracy of responses was unrelated to condition, which indicates that neither condition consistently produced more informative exchanges than the other. This is particularly apparent in the Round 1 response data, where the cooperative condition produced longer exchanges on average, yet they have a lower accuracy than the competitive condition. Such a data pattern could suggest that shorter exchanges are more informative – that is, perhaps the shorter exchanges are more specific and easy from which to accurately determine the referenced tangram – but this supposition is not borne out by the fluctuating accuracy of the cooperative condition.

Informativity is thus not a viable explanation for the difference in exchange length between conditions; the cooperative condition’s longer exchanges are not more informative – more helpful to another person – than the competitive condition’s. The condition by round interaction that was found is unrelated to duration or word count; neither predicted accuracy of selecting the referenced tangram. This interaction is a result of the accuracy of responses to the
first round of the cooperatively-produced exchanges compared to later rounds’ exchanges. One explanation for this result could be that the first-round tangrams were easier to differentiate with descriptions than later rounds’ tangrams were, and so the competitive players could describe the tangrams in the first round better than in later rounds. Another possible explanation is that the cooperative players formed conceptual pacts (Brennan & Clark, 1996) that made it more difficult for “overhearers” to understand their descriptions in Rounds 1, 2, and 4, which were the rounds in which the cooperatively-produced exchanges were understood less well than the competitively-produced exchanges (Schober & Clark, 1989).

Experiment 1b did not address the question of why the conditions behaved very differently at the beginning of the five rounds, and at the first instance of a given tangram exchange, but approached the same length of exchanges later. Given that the players’ goals appeared to influence their formation of common ground, the players’ maintaining their goals may have led to goal habituation. Experiment 2 addressed the question of whether or not this occurred.
Chapter 4: Experiment 2

In Experiment 1a, the difference between the average exchange lengths of competitive pairs and cooperative pairs decreased over the rounds of the game. This could be due to goal habituation: it is cognitively demanding to maintain the goal of being cooperative or competitive throughout the game (Ariga & Lleras, 2011). Ariga and Lleras (2011) show that taking brief breaks from maintaining a task goal keeps people from slipping. If goal habituation is occurring, then this would provide evidence for the presence of extra-linguistic goals that affect reference production.

There are two goals for Experiment 2: to replicate the findings of Experiment 1a and examine whether goal habituation explained the pattern of results in Experiment 1a. If the players become habituated to their goals over the rounds of the experiment, they would be less competitive or cooperative than they originally were. Specifically, the pairs’ exchange lengths will resemble the start-of-game lengths after breaks in the game, after the goals are re-sensitized; cooperative players’ exchanges should lengthen after a break, and competitive players’ exchanges should shorten after a break.

To meet these goals, we specifically tested the performance of cooperative pairs against competitive pairs in a game very similar to Experiment 1a. Crucially, Experiment 2’s setup was a close variant of the original game, to act as a replication experiment; however, there were breaks between certain rounds in the games, so that the players would become re-sensitized to their goals (Ariga & Lleras, 2011). As a consequence of needing to compare before and after the breaks, there were more rounds in the game. Experiment 2 will thus test in another way whether goals affect how interlocutors come to refer to items under discussion.
**Method**

**Participants**

A total of 108 participants in pairs did the experiment for either course credit (41 pairs, all undergraduates at the University of Illinois at Urbana-Champaign) or for $16 (23 pairs). All were native English speakers and were naïve to the experimental manipulation. Eight pairs replaced because at least one participant was not a native English speaker, and six pairs were replaced because they failed to follow instructions.

**Procedure**

The pairs of participants played a version of the game used in Experiment 1a. This version was played on two separate computers in the same room, so that participants could hear each other but not see each other’s screen. They were randomly assigned the role of Player 1 or Player 2. The experimenter explained the game just as in Experiment 1a, including an example of the question-answer-clarification format. The only differences were that a physical demonstration was not employed and that a ship could not be located on the other player’s screen where one of their own was located.

In each round, the players viewed the same five tangram images in a horizontal row, but in different orders on their individual computer screens. Each tangram image was surrounded by four small square buttons, one above, below, to the right, and to the left; players were informed that these were the “ships.”

The players took turns asking for the location of the other player’s ships around the tangram images and answering the other player’s questions. They were instructed to stick to a question-answer-clarification format. Each player knew where her ships were located, but not the other player’s ships. Players could describe the images whatever way they wished, but were
instructed to not use locative phrases (e.g. the picture on the right) because these phrases would indicate different pictures for each of them. The players’ voices were recorded directly to disk. The experiment took about 80 minutes.

Pairs of participants were randomly assigned to play either competitively or cooperatively. Like Experiment 1a, the game was identical in both conditions: find all four of one player’s ships. This game was presented under different goals, playing competitively to find all four of the other player’s ships to win the round, or playing cooperatively to find all four of one player’s ships to end the round together.

To play the game, players clicked on the ship they were guessing about, asked their question and got a response, clicked again if they guessed incorrectly, and clicked once more on that ship if the other player did have a ship located there. Each click on a ship was recorded.

Below the tangrams and ships on the screen were a “Next” button and instructions to click the “Next” button when all four of one player’s ships were found. When the “Next” button was clicked by a player, she saw a message that reinforced the goal. For cooperative players, the message was, “Congrats for completing the round together.” For competitive players, if a player indicated through appropriate clicks that she had found the other’s four ships, the message “Congrats for winning the round!” appeared; if she did not, the message “Sorry, you did not win” appeared. The players then were instructed who would ask first in the next round. (Who would ask first was counterbalanced across conditions and tangram set order.)

After Rounds 3 and 6, players saw a screen instructing them to leave the room. They came out and sat with the experimenter in another room and were instructed to take a break; no restrictions on break activities were given. After about ten minutes, both players were escorted back into the experiment room and instructed to resume play, with a verbal reiteration of their
prior competitive or cooperative instructions for gameplay. This manipulation tested whether their performance is affected by re-sensitizing them to their extra-linguistic goals.

The players finished the game after completing nine rounds; then they were debriefed.

_Materials_

Each participant completed nine rounds; each round included five new tangrams. The tangrams were humanoid figures of varying degrees of difficulty. The order of tangram presentation on the screen was different within pairs, such that (for example) Player 1 saw the tangrams in the order A B C D E and Player 2 saw the tangrams in the order E A D C B. There were four set orders: sets 1-9 (for Rounds 1-9), sets 1-9 randomly reordered, sets 10-18 (for which all 45 tangrams were randomly reshuffled into new sets), and sets 10-18 randomly reordered. These set orders were counterbalanced across conditions.

The ship locations were pseudorandomly assigned such that no two ships were ever placed at the same location within a round. This was done to eliminate a recurring source of confusion for players in Experiment 1a. Some players appeared to forget that they could guess the locations on the other player’s board that were where their own ships were located on their own board, because they did not ask about those locations until no other possible locations were available. This confusion also manifested in not realizing that two markers (a ship and a guess) could be located in the same position. In Experiment 2, this source of confusion was eliminated. (Note: this decision did result in some players’ employing the strategy of not guessing the places that the other player asked about, because those places could not be ship locations for the other player.) The order of the ship locations was reversed across rounds for half of the players.
Predictions

The goal habituation hypothesis predicts that each condition’s behavior will begin very differently, then gradually become more similar, but then shift back to being very different after the condition’s goal is re-sensitized from the break. Experiment 1a suggests that competitive players begin with shorter exchanges, and cooperative players begin with longer exchanges. If their gradual convergence on exchange length is due to goal habituation, then I expect that, after a break, the competitive players will produce shorter exchanges and the cooperative players will produce longer exchanges. They will then habituate to their goals and become more similar in exchange length until after the next break. The pre- and post-break exchange lengths must be compared. To replicate Experiment 1a, the word count and duration of the exchanges must be significantly different between conditions, both across and within rounds of the game.

Results

The word count and duration of the exchanges were compared as before. I and five undergraduate assistants, naïve to the experimental conditions of given pairs, transcribed the conversations of the players. The exchanges were coded as before: all words pertaining to the tangram description were counted, excepting filler words and disfluencies. As in Experiment 1a, mistakes – in which a player asks to discuss a misunderstanding about which terms belong to which tangram – were included in the analysis, as there were only eleven. Phrases such as the same one were discarded from analysis, as these simply callbacks to the immediately-previous exchange, rather than exchanges establishing a reference; locatives were also discarded from analysis, as they were not actual exchanges establishing a reference and were against instructions; this eliminated only 8% of the exchanges.
Duration

Similarly, because the length of the entire exchange was of interest, an undergraduate assistant and I subsequently marked the duration of the exchanges about each tangram, including all pauses during the exchange. Comparison of 2699 exchanges suggested that there were no clear differences in the durations between conditions, nor were there differences pre/post break. The average duration of Competitive exchanges was 8.28 seconds and the average duration of Cooperative exchanges was 8.77 seconds. Across rounds, the average duration for competitive players in the first round was 13.80 seconds, decreasing to 6.62 seconds in round nine, which is in the opposite direction from what the Experiment 1a results predicted. The lack of differences between the conditions meant further coding of the durations was deemed unnecessary; the durations are plotted in Figure 6 for comparison.

![Figure 6](image)

**Figure 6.** Average duration over rounds by condition, for a subset of the Experiment 2 data (15 pairs). Competitive pairs’ exchange lengths decrease over time, opposite to prediction.

Word Counts

The word counts of the exchanges were compared across rounds between conditions (see Figure 7). The average exchange length of the Competitive condition was 9.33 words, and 8.72
words for the Cooperative condition. The conditions were compared across rounds using a mixed effect model with subject and pair as random effects, and condition and round as fixed effects; the results are listed in Table 4.

![Graph showing average word count of tangram exchanges across instances.](image)

**Figure 7.** Average word count of tangram exchanges across instances, using all rounds and tangrams, compared between conditions. Blue brackets mark when the breaks occurred: between Rounds 3 and 4 and Rounds 6 and 7.

**Table 4**
Parameter estimates by subjects, tangrams, and pairs for analysis of condition and round effects in word count. The data analyzed included 11358 exchanges from 128 subjects (64 pairs) and 45 tangrams.

|                  | Estimate | SE  | z-value | Pr(>|z|) |
|------------------|----------|-----|---------|----------|
| Fixed effects    |          |     |         |          |
| Intercept        | 2.15     | 0.03| 66.34   | .185     |
| Condition        | -0.09    | 0.06| -1.33   | .185     |
| Round            | -0.02    | 0.01| -2.56   | .010*    |
| Condition*Round  | -0.01    | 0.02| -0.39   | .696     |
|                  |          |     |         |          |
| Variance         |          |     |         |          |
| Std. Dev         |          |     |         |          |

<table>
<thead>
<tr>
<th>Groups</th>
<th>Variance</th>
<th>Std. Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject (Intercept)</td>
<td>0.02</td>
<td>0.14</td>
</tr>
<tr>
<td>Subject Round</td>
<td>0.001</td>
<td>0.03</td>
</tr>
<tr>
<td>Pair (Intercept)</td>
<td>0.06</td>
<td>0.24</td>
</tr>
<tr>
<td>Pair Round</td>
<td>0.003</td>
<td>0.06</td>
</tr>
</tbody>
</table>

*p<.05
This shows a significant effect of round ($z = -2.26, p<.05$). However, there is no effect of condition ($z = -1.33, p>.05$) and no interaction between condition and round ($z = -0.39, p>.05$). Analysis of the simple slopes shows that the word count of the cooperative exchanges significantly decreases over the nine rounds ($z = -2.10, p=.035$), but the competitive exchange lengths do not differ from zero ($z = -1.51, p=.131$). The decrease in length in the cooperative exchanges could drive the across-rounds differences. Like for duration, the average word counts do not match the pattern of the Experiment 1a findings in the competitive condition.

As in Experiment 1a, the number of exchanges that included a clarification about the terms used was compared across rounds between conditions. The model used a binomial distribution – whether there was a clarification or not. There was an effect of round ($z = 3.28, p<.05$) and condition ($z = 3.02, p<.05$), but no interaction. This reflects the 961 exchanges with clarifications in the competitive condition compared to 722 such cooperative exchanges.

The exchange lengths pre- and post-breaks were compared using only the rounds occurring before or after a break, which occurred between Rounds 3 and 4 and between Rounds 6 and 7. To determine whether the performance over the entire game differed, Early rounds (3 and 4) were compared with Late rounds (6 and 7). To determine whether the goals were re-sensitized, the performances were compared between Pre-break rounds (3 and 6) and Post-break rounds (4 and 7). There was a main effect of Pre/Post ($z = -2.05, p<.05$), Early/Late ($z = 2.40, p<.05$), and an interaction between them ($z = -3.50, p<.001$). The interaction is driven by the Late rounds ($z = -6.03, p<.001$). However, visual inspection of Figure 7 shows that the conditions behave in opposite patterns from Early to Late: the cooperative decreases from Round 3 to 4 but increases from Round 6 to 7, and vice versa for competitive.
Next, the conditions’ performances within rounds were compared to determine whether there were any differences between the word counts of repeated references, which would illustrate the process of forming common ground on the referenced tangrams (see Figure 8). The model was the fullest that converged, following the procedure described by Barr, Levy, Sheepers, & Tily (2013). The model contained subject, tangram, and pair as random effects, with instance, round, and condition as fixed effects. The word counts within rounds were compared between conditions using a Poisson distribution, using a subset of the previously-analyzed data. This subset is comprised of all rounds in which players correctly followed instructions for marking the tangrams guessed, comprising 332 of 576 possible rounds.

The best-fit model indicated that the interaction between condition and instance was significant, \( z = 3.20, p < .05 \). The effect of condition on instance was explored with the first, second, and then all other instances; there was a significant effect of condition on the first instance \( z = -2.18, p < .05 \) but not the others, suggesting that the first instance drives this interaction. No other interactions were reliable. There is a main effect of instance \( z = -17.53, p < .001 \), which suggests that players’ exchanges shortened over time, though may be an artifact of the imbalance of the number of instances (there are more first instances than any other instance). There is a main effect of condition \( z = -2.51, p < .05 \) and of round \( z = -1.99, p < .05 \). The model output is displayed in Table 5.
Figure 8. Average word count of tangram exchanges across instances compared between conditions. The data analyzed included 6775 exchanges.

Table 5
Parameter estimates by subjects, tangrams, and pairs for analysis of condition, round, and instance effects in word count. The data analyzed included 6775 exchanges from 111 subjects (56 pairs) and 45 tangrams.

|                               | Estimate | SE  | z-value | Pr(>|z|) |
|--------------------------------|----------|-----|---------|----------|
| **Fixed effects**              |          |     |         |          |
| (Intercept)                    | 2.76     | 0.07| 40.21   | .012*    |
| Condition                      | -0.33    | 0.13| -2.51   | .012*    |
| Round                          | -0.02    | 0.01| -1.99   | .046*    |
| Instance                       | -0.26    | 0.01| -17.53  | <.001*   |
| Condition*Round                | 0.002    | 0.02| 0.08    | .937     |
| Condition*Instance             | 0.09     | 0.03| 3.20    | .001*    |
| Round*Instance                 | 0.001    | 0.003| 0.31   | .757     |
| Condition*Round*Instance       | -0.002   | 0.007| -0.34  | .732     |

<table>
<thead>
<tr>
<th></th>
<th>Variance</th>
<th>Std. Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Random effects</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Groups</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subject (Intercept)</td>
<td>0.062</td>
<td>0.25</td>
</tr>
<tr>
<td>Instance</td>
<td>0.004</td>
<td>0.06</td>
</tr>
<tr>
<td>Round</td>
<td>0.005</td>
<td>0.07</td>
</tr>
<tr>
<td>Instance*Rd</td>
<td>0.0003</td>
<td>0.02</td>
</tr>
<tr>
<td>Tangram (Intercept)</td>
<td>0.017</td>
<td>0.13</td>
</tr>
<tr>
<td>Pair (Intercept)</td>
<td>0.201</td>
<td>0.45</td>
</tr>
<tr>
<td>Instance</td>
<td>0.009</td>
<td>0.10</td>
</tr>
<tr>
<td>Round</td>
<td>0.003</td>
<td>0.05</td>
</tr>
<tr>
<td>Instance*Rd</td>
<td>0.0003</td>
<td>0.02</td>
</tr>
</tbody>
</table>

*p<.05
Discussion

The patterns of forming shared references found in Experiment 1a were not matched in Experiment 2. The competitive condition did not significantly increase the word count of the exchanges over time, and there was no effect of condition on the length of the exchanges. The goal habituation hypothesis predicted that the competitive condition would decrease the word count after each goal-resensitizing break, but this did not occur. Neither did the cooperative condition increase the word count after each break. The Late break did have an effect, but the cooperative and competitive conditions acted oppositely from the previous findings.

The number of clarifications was greater for the cooperative condition in Experiment 1a, but was greater for the competitive condition in Experiment 2. In Experiment 2, the competitive condition had a greater average word count (9.33), which matches the greater number of clarifications. This suggests that perhaps the back-and-forth exchanges included more description of the tangrams.

From the subset analyzed with instance as a factor and tangram as a random intercept, it appears that, within rounds, there are effects of condition, round, and instance on word count, and an interaction between instance and condition; Experiment 1a also had effects of instance and a condition by round interaction. This suggests that the rate of shortening the exchanges was different by condition within rounds in each experiment, despite the differences in the patterns of exchange lengths across rounds. This likely reinforces the universality of shortening reference productions with repeated references rather than suggesting anything different between the conditions’ performances.
Chapter 5: General Discussion

The process of establishing shared references for the tangrams under discussion appeared to follow the informativity hypothesis in Experiment 1a: cooperative players shared more information, and competitive players shared less. However, the exchanges used in Experiment 1a were not differently informative between conditions when tested on overhearers for accuracy in selecting the tangram discussed. Furthermore, the pattern of reference productions differed between Experiments 1a and 2; the conditions had inconsistent effects.

One explanation for these inconsistent findings is that the task in Experiment 1a was slightly different from the task in Experiment 2. There were two main non-experimental-design differences between the tasks. The first was that the players in Experiment 1a were being observed by the experimenter during the task, whereas the experimenter did not observe participants in Experiment 2 during the experimental task. This might have influenced participants in Experiment 1a to meet experimenter expectations, nebulous though they were. Competitive players might have naively assumed that saying less would be less helpful to the other person, for instance. Or, as Kuhlen & Brennan (2013) suggest about confederates in language tasks, perhaps the experimenter’s presence led to the players behaving in inauthentic or unusual ways.

In Experiment 2, the lack of a co-present experimenter may have led to more confusion about the task in general, resulting in unexpected conversational patterns. Because the experimenter was in a separate room, assistance in understanding the game was less accessible. As a consequence of the players in both conditions being confused by the game, they would produce exchanges more influenced by confusion with how to play the game than by condition. Thus, both conditions would show similar performance.
The second main difference between the experimental tasks was that Experiment 2 utilized a computerized version of the game. This version of the game was less like the actual game of Battleship than Experiment 1a, because of the constraints of playing the game on computers. If players in both conditions were focused on the experience of navigating the game on the screen, the exchanges would be affected less by condition than by the perceived strangeness of the task. This would lead to similar performance between conditions, as was found. The constraints of computerizing the game may have influenced players’ expectations about what they were supposed to do during the experiment, making everyone’s extra-linguistic goal the same – figure out what the experiment was about – rather than different.

Yet another explanation is that the experimental manipulation was inconsistently successful. That is, perhaps the way of establishing cooperation or competition worked for some pairs but not all of them. It is difficult to know how differences could exist between conditions in Experiment 1a if the results were random noise, but the patterns of Experiment 2 – a much larger dataset – may bear out the assertion that Experiment 1a’s results were indeed chance.

Further experimentation could attempt to establish extra-linguistic goals more firmly and realistically, perhaps having the pairs play a separate game first in which the strategies for competitive versus cooperative players would be very distinct, thus clearly guiding them into the desired goal state. There seemed to be an effect of the breaks, lending some credence to the value of breaks when maintaining a cognitive goal (Ariga & Lleras, 2011); the effect of breaks on linguistic cooperation could be fruitful to study. For example, the effect of taking breaks when establishing shared references with people of different expertise levels or with people of different social or cultural backgrounds (Fussell & Kraus, 1992).
Chapter 6: Conclusions

The aim of this work was to determine whether extra-linguistic factors, specifically goal states, affect how interlocutors establish referring expressions for the same discourse entity. Interlocutors may linguistically align with each other in a conversation (Pickering & Garrod, 2004), but their goals may not be the same. Since the extra-linguistic aspects of a conversation can influence interlocutors’ linguistic choices (Isaacs & Clark, 1987; Russell & Schober, 1999; Wardrow Lane & Liersch, 2011), I examined whether task goals in a game would influence the referring expressions developed by competitive and cooperative players. The present research addressed this by examining how the length of exchanges about shared items under discussion differed between competitive and cooperative conditions. The exchanges were investigated because, as interlocutors come to share the terms they use for an item under discussion, their referring expressions for that item grow shorter (Clark & Wilkes-Gibbs, 1986).

Experiment 1a found a condition by round effect for the exchange length, particularly driven by the first instance of referring to a tangram. This suggests perhaps the cooperative players were over-informative in the Gricean (1975) sense, or that the competitive players were under-informative. Both conditions shortened their exchanges within and across rounds, just as Clark and Wilkes-Gibbs (1986) would predict: the interlocutors established more efficient, shared terms over time.

To determine whether there were any differences in informativity in these exchanges produced, Experiment 1b tested the accuracy of overhearers at selecting the correct tangram after hearing the original exchange. The exchange lengths were set as predictors of informativity – however, exchange length and accuracy were unrelated. The informativity hypothesis was not
supported. This suggested that the competitive condition was not simply being under-informative and difficult to understand compared to the cooperative condition.

Experiment 2 used a theory of the nature of cognitive goal maintenance to attempt to demonstrate how goals affect linguistic choices. Ariga and Lleras (2011) showed that taking breaks can re-activate a waning goal representation. Habituating to a goal – getting used to a goal and having difficulty maintaining it – was a possible explanation for the pattern of exchange lengths found in Experiment 1a, and finding an effect of goal habituation on exchange lengths would suggest further that extra-linguistic goals influence how interlocutors establish shared terms. However, the findings instead did not consistently resemble the patterns of exchange lengths of Experiment 1a: for example, the cooperative players started the game with shorter exchanges than the competitive players. The first instance of referring to a tangram again drove a condition by instance interaction, and there was a main effect of instance, meaning that pairs in both conditions shortened their exchanges over time. The predicted effect of goals on exchange lengths was not found in Experiment 2. Although both conditions changed their exchange lengths after breaks, these changes were either positive then negative or vice versa, so, the breaks in Experiment 2 had inconsistent effects on the exchange lengths.

Despite the inconsistency of the effects of condition on exchange lengths, there were observed consistent effects of round and instance on exchange lengths. These findings fit with the larger literature on the formation of conceptual pacts (Brennan & Clark, 1996): as tangrams were repeatedly referenced, the exchanges establishing reference shortened. As the players became more expert at deciphering and describing the tangram shapes during the game, and communicated with another burgeoning tangram expert, their exchanges gradually shortened as well (Isaacs & Clark, 1987; Russell & Schober, 1999).
Goals are important in conversation; there are clearly different linguistic choices made when one is fighting compared to when one is comforting another person, for example (Sarangi & Slembrouck, 1992). However, across both games played in these experiments, consistent effects of goals were not observed.
References


