Dialogue & Discourse 12(1) 45-72 doi: 10.5210/dad.2021.103

Referential Communication Between Friends and Strangers in the Wild

Kris Liu KYLIU@UCSC.EDU
University of California, Santa Cruz

J. Trevor D'Arcey JDARCEY@UCSC.EDU
University of California, Santa Cruz

Marilyn Walker MAWALKER@UCSC.EDU
University of California, Santa Cruz

Jean E. Fox Tree FOXTREE@UCSC.EDU
University of California, Santa Cruz

Editor: Barbara Di Eugenio

Submitted 02/2018; Accepted 12/2020; Published online 04/2021

Abstract

The Map Task (Anderson et al., 1991) and Tangram Task (Clark & Wilkes-Gibbs, 1986) are traditional referential communication tasks that are used in psycholinguistics research to demonstrate how conversational partners mutually agree on descriptions (or *referring expressions*) for landmarks or unusual target objects. These highly controlled, laboratory-based tasks take place under conditions that are relatively unusual for naturally-occurring conversations (Speed, Wnuk, & Majid, 2016). Using the Artwalk Task (Liu, Fox Tree, & Walker, 2016) – a real-world-situated blend of the Map Task and Tangram Task – we showed that the process of negotiating referring expressions "in the wild" is similar to the process that takes place in a laboratory. In addition to replicating laboratory results showing lexical entrainment, we also found that acquaintanceship and extraversion influenced the number of unique descriptors used by pairs. In round 1, introverts in stranger pairs used fewer descriptors but introverts in friend pairs were indistinguishable from extraverts. The influence of extraversion declined by round 2. Lexical entrainment observed in labs is generalizable to real-world settings, and lexical entrainment in naturalistic communication, at least, is subject to social and personality factors.

Keywords: lexical entrainment, acquaintanceship, referential communication tasks, outdoor tasks, extraversion

1 Introduction

Many laboratory-based tasks have characteristics that are not found in naturalistic settings. They take place in relatively sterile environments where distractions are discouraged, and participants are frequently separated from other people who are not involved in the experiment. Participants are often seated at computers in booths, a situation that more closely resembles an exam than a conversation. While participant pairs engaging in lab-based discourse tasks may occasionally wander off-topic in their talk, there is usually little to tempt them to do so. In contrast to laboratory communication tasks, many real-world conversations take place amid noise, people, and other distractions. The following excerpt from the current study shows the sort of side conversation that is unlikely to happen without external distractions: a *matcher* who is on a mobile phone on a public

street is receiving instructions from a remote *director* about the location and description of a piece of public art (in this and later examples, D refers to the director and M refers to the matcher):

(1) M: canvassers are everywhere
D: did he hassle you for being on the phone or something?
M: [laugh] yeah like this guy came up to me and like tried to run in front of me
D: what?
M: I was just like, "I'm doing a psychology experiment" and he was like, "Oh you're doing the fake phone thing?" and I was like, "What? I'm on the phone" [pause] I could never do that to someone

Our investigation shows that lexical entrainment, a reliable finding in laboratory settings, also occurs in more naturalistic settings. In the following example, the director proposes the word *blob*, among other descriptions, to describe an art object:

(2) D: okay so we're looking for uh this like concrete thing I don't know how to explain it
M: it's a concrete thing?
D: yeah okay
M: um is it is it a sculpture?
D: uh yes it's like a big just blob I couldn't really describe it to yo- it looks almost like a jellyfish

When the matcher finds the object, the matcher shows acceptance of the label *blob* by saying, "I see the blob I see the blob." Later in the experiment, the director instructed the matcher to find the object again with the instructions, "Do you remember the big blob?" which is a shorter version of the original description of a blob that was a "concrete thing" that "looks almost like a jellyfish." The matcher entrained on the label *blob* saying, "Okay I'm taking a picture of the blob right now."

The above examples took place during a *referential communication* task, a traditional experimental paradigm that is used to study collaborative communication. In one referential communication task, the Map Task (Anderson et al., 1991), directors and matchers are given two different maps: Maps share some features (but not all), and the director's map has a specific route drawn on it whereas the matcher's map has no route. It is the director's job to verbally guide the matcher to draw the director's precise route through the landmark without actually showing the matcher the director's own map. The Tangram Task has a similar asymmetric setup: The director has a set of tangrams (or abstract shapes) in a specific order and the matcher has access to the same tangrams (plus some distractor tangrams). The director must verbally guide the matcher to identify the same tangrams the director sees and get the matcher to place them in the same order as they are shown in the director's set (Clark & Wilkes-Gibbs, 1986; Schober & Clark, 1989). The task is often done more than once in an experimental session, allowing researchers to observe how collaborative behavior changes across iterations or rounds. Both experimental paradigms clearly show that directors and matchers work toward common understanding (*grounding*, e.g., Clark, 1996) including coming to use (and re-use) the same terms for objects shared between their asymmetric perspectives (*entrainment*, a term used here mostly synonymously with Clark's notion of collaboration on referential language).

We developed a naturalistic task that is conceptually similar to the wayfinding Map Task and object-identifying Tangram Task but also situates the matcher in a walkable downtown area while the director directs the matcher to a series of public artwork *targets* via a Skype-to-mobile phone connection. By taking the matcher out of the lab and making the director contend with a matcher who is in a complex environment, the task more closely resembles the natural setting of many conversations, particularly those conducted with mobile devices. The mobile demands of the task require matchers to tune out surrounding noise and conversation. They need to consider street crossings (Neider, McCarley, Crowell, Kaczmarski, & Kramer, 2010) as well as social constraints. For instance, talking on a mobile phone in public for an extended amount of time can be viewed as being exasperating or impolite (Love & Perry, 2004). Attending to social concerns while moving about a city can reduce attention to a mobile phone task (Oulasvirta, Tamminen, Roto, &

Kuorelahti, 2005). In addition, walking can negatively impact cognitive performance (Beauchet, Dubost, Herrmann, & Kressig, 2005; Hill, Bohil, Lewis, & Neider, 2013). And conversely, engaging in cognitive tasks can slow down walking or change a person's gait (Yogev-Seligmann, Hausdorff, & Giladi, 2008).

The presence of bystanders may also influence participants' behavior. In laboratories, participants may be observed by an experimenter or they may be left alone in a room or booth; in the case of Artwalk, only directors were sitting alone in a booth – matchers were out on a busy street, walking or standing near a variable number of bystanders. Asking for directions on a mobile phone may not be unusual but attempting to describe an abstract piece of artwork to someone who seems to require multiple different descriptions, snapping a photo, and then repeating the process with another piece of art is not typical street behavior. This has the potential to exacerbate participant self-consciousness, which can influence behavior (Froming & Carver, 1981).

Nevertheless, it should be noted that the relationship between laboratory and non-laboratory behavior can be hard to predict. For example, laboratories may seem to naturally engender more formal or deliberate speech than non-laboratory settings but not all experimental linguistic work has supported this (cf. Xu, 2010). In our own laboratory, deliberate effort by experimenters was required to create a formal enough atmosphere to observe changes in speakers' uses of the quotation devices *said* and *like* (Blackwell & Fox Tree, 2012).

Our first goal with the Artwalk Task (Liu, Fox Tree, & Walker, 2016) was to assess whether entrainment patterns (mutual adoption of common terms, as measured by their repeated use) observed in laboratories could be replicated in a more naturalistic setting involving one mobile communicator and one non-mobile communicator. It is possible that the distractions, social concerns, and different pacing of the task in a naturalistic setting could affect lexical entrainment. Our second goal was to test whether we could observe differences in communication based on two social factors: acquaintanceship status and extraversion. Acquaintanceship was part of the original design of the Map Task (Anderson et al., 1991), although few acquaintanceship differences were documented (see Acquaintanceship and Communication section). Extraversion is not often studied in the context of referential communication (see Extraversion in Collaboration and Conversation section). We hypothesized that both acquaintanceship and extraversion would affect communication in this naturalistic setting. With this novel, real-world task we reproduced effects observed in laboratories and additionally demonstrated that some interpersonal and personality variables can affect how efficient pairs are at completing a referential communication task.

1.1 Referential Communication in the Laboratory

Referential communication tasks in the laboratory have yielded consistent results on the process of coordinating referring expressions between conversational partners. When two conversational partners talk about objects that are not easily named, they undergo a period of negotiation where they have to ensure that they are both discussing the same object (grounding) and then implicitly or explicitly agree on what to call it (entrainment). Over time, referring expressions shorten and communication becomes more efficient (Clark & Wilkes-Gibbs, 1986). For example, a participant may initially identify a specific tangram as "All right, the next one looks like a person who's ice skating, except they're sticking two arms out in front" but later refer to it as "the ice skater" (Clark & Wilkes-Gibbs, 1986, p. 12). Participants entrain on the labels used to refer to objects, such as observed with the use of the label *blob* in Example 2.

Many factors can influence grounding, including the level of expertise with the topic under discussion (Isaacs & Clark, 1987), the level of expertise with the communicative medium (Fox Tree, Mayer, & Betts, 2011), culture (Wang, Fussell, & Setlock, 2009; Wu & Keysar, 2007), and age (Horton & Spieler, 2007; Kemper, Othick, Warren, Gubarchuk, & Gerhing, 1996). In this study, we examine the influence of acquaintanceship and extraversion.

1.2 Acquaintanceship and Communication

Whether or not interlocutors know each other changes communicative behavior. In some previous studies investigating the differences between friends and strangers, friends outperformed strangers. Friends were better at guessing the thoughts and feelings of friends (Stinson & Ickes, 1992) and they more easily identified a target figure when described by a friend than by a stranger (Fussell & Krauss, 1989). Friends were also better at sending covert messages to each other without the covert message being understood by strangers (Fleming, Darley, Hilton, & Kojetin, 1990). Retrieval cues generated by friends were more helpful for word recall than those generated by strangers (Andersson & Rönnberg, 1995).

In other previous studies investigating the differences between friends and strangers, being friends was found to negatively impact performance. Friends tended to project their own knowledge and beliefs on each other, resulting in incorrect assumptions of clarity (Savitsky, Keysar, Epley, Carter, & Swanson, 2011). They were also more likely to overestimate the success of their performance as a pair in comparison to strangers (Gould, Osborn, Krein, & Mortenson, 2002). In recall tasks, friends were more likely to induce false memories in each other (Hope, Ost, Gabbert, Healey, & Lenton, 2008).

Beyond performance, the structure of discourse was also found to differ between friends and strangers. Friends were more likely to use informal language, to interrupt, to talk about multiple topics, to self-disclose, to be judgmental, and to ask for favors (Planalp & Benson, 1992). Friends also used more discourse markers (Fox Tree, 2007; Jucker & Smith, 1998), overlapped more turns (Bortfeld, Leon, Bloom, Schober, & Brennan, 2001; Dunne & Ng, 1994), laughed more (Smoski & Bachoroski, 2003), and left more information unspoken in openings and had more complex closings (Hornstein, 1985).

Yet some researchers found no discernible effect of acquaintanceship on some discourse phenomena, including no effect on disfluencies (Bard, Aylett, & Lickley, 2002; Branigan, Lickley, & McKelvie, 1999), little effect on prosodic convergence (Truong & Heylen, 2012), and no effect on the amount of laughter (Truong & Trouvain, 2012). The inconsistent effect of acquaintanceship on discourse may account for why few researchers, after thirty years and hundreds of citations, have documented differences between friend and stranger pairs in the Map Task Corpus, although an even split between friend and stranger pairs was part of the original design of the study (Anderson et al., 1991).

The lack of differences between friends and strangers in the Map Task Corpus suggests that acquaintanceship will not affect grounding and entrainment behaviors with the Artwalk Task. Nonetheless, acquaintanceship differences may be more apparent in a more natural setting. If we do find differences based on acquaintanceship in the wild, this raises the question of whether lab settings dampen or extinguish acquaintanceship differences. Lab settings may have that effect because they focus participants' attention on the tasks at hand. With the Artwalk Task, participants often talked freely about off-task topics as they walked between art objects (Guydish, D'Arcey, & Fox Tree, 2020).

1.3 Extraversion in Collaboration and Conversation

Extraversion is one of the most readily and reliably recognizable personality traits at zero acquaintance (Albright, Cohen, Malloy, Christ, & Bromgard, 2004; Kenny & Acitelli, 2001; Levesque & Kenny, 1993) with social and linguistic behaviors that are identifiable even when not in face-to-face interaction (Gill & Oberlander, 2003). Although level of extraversion is a continuous scale, in the following discussion, we refer to extraverts and introverts as shorthand for those who score higher or lower on the scale.

Extraverts have an advantage for some types of laboratory studies, particularly those that involve greater external pressure. They performed better than introverts on verbal working memory tasks that had time limits (Rawlings & Carnie, 1989). They also performed better on both practiced

and novel tasks when there was someone observing them (Uziel, 2007). In groups, extraverts generated more unique and diverse ideas than introverts (Jung, Lee, & Karsten, 2012), although the presence of highly extraverted participants on teams did not always translate to better team performance (Peeters, van Tuijl, Rutte, & Reymen, 2006).

In discourse, extraverts tended to take the lead in conversations (Cuperman & Ickes, 2009). They said more, spoke louder, spoke faster, used fewer pauses, and used more backchannels (Campbell & Rushton, 1978; Carment, Miles, & Cervin, 1965; Dewaele & Furnham, 1999; Feldstein & Sloan, 2984; Gifford & Hine, 1994; Scherer, 1978). Extraverts' speech had lower lexical richness and used a greater proportion of verbs, adverbs, and pronouns (Dewaele & Furnham, 1999). Introverts tended to choose their words more carefully, qualifying their statements more frequently (Oberlander & Gill, 2006; Pennebaker & King, 1999). Introverts also used more descriptive and concrete language (Beukeboom, Tanis, & Vermeulen, 2013), although extraverts used more adjectives than introverts (Gill & Oberlander, 2003). In casual conversations and in formal examinations, extraverts relied on greater shared knowledge between themselves and their interlocutors (Heylighen & Dewaele, 1999). They were more likely to use deictic expressions (e.g., pronouns such as *he*) than introverts, whose language tended to be more explicit about referents (Heylighen & Dewaele, 1999). In other words, extraverts were more likely than introverts to tailor their speech to the context of a conversation. This tailoring can also be observed with spontaneous gestures that extraverts produce (Tolins, Liu, Neff, Walker, & Fox Tree, 2016).

Prior work on language differences between extraverts and introverts make no clear predictions about how grounding and entrainment behaviors differ based on this personality variable. Nonetheless, extraversion differences, if they exist, may be more likely to manifest in the naturalistic Artwalk Task where participants had more opportunity to engage in off-task dialogue.

1.4 Acquaintanceship and Extraversion

Personality and acquaintanceship can interact in discourse. In one study, extraverts were better at conducting conversations with strangers (Thorne, 1987), possibly due to their increased willingness to take the lead (Cuperman & Ickes, 2009). But within friend pairs, introverts were quite assertive, particularly when both friends were introverts (Nelson, Thorne, & Shapiro, 2011). In mixed pairs, however, only 40% of introverts were reported as being as engaged in speaking as their extravert friends (Nelson et al., 2011).

As was observed in the laboratory, acquaintanceship and extraversion may also interact in our study. For example, personality variables may be more pronounced in stranger pairs than in friend pairs.

1.5 Current Study

We tested the role of acquaintanceship status and extraversion on communicative efficiency during a real-world-situated referential communication task. Communicative efficiency is an observable dimension of entrainment and was defined as the shortening of referential descriptions. Shortening of referential descriptions is a linguistic behavior that has been shown repeatedly in laboratory referential communication tasks (Clark & Wilkes-Gibbs, 1986; Schober & Clark, 1989; Tolins, Zeamer, & Fox Tree, 2018) as well as in an outdoor task that took place in a controlled experimental setting (Brennan, Schuhmann, & Batres, 2013). We predicted that shortening would also occur in a far more complex, real-world referential communication task conducted in distracting communicative circumstances.

A traditional assessment of communicative effectiveness in referential communication is the Tangram Task. This task involves assessing the number of correct identifications from a dozen identifications (Clark & Wilkes-Gibbs, 1986), often from a larger array (Fox Tree, 1999; Fox Tree & Clark, 2013), with a game that is played repeatedly, such as with eight iterations (Tolins et al., 2018). In the traditional task, directors and matchers see all the two-dimensional items-to-be-placed

at the same time, often on a table in front of them or on a computer screen. In the Artwalk Task, directors saw items-to-be-placed one at a time and matchers saw multiple pieces of public art as they walked around the downtown area. Because the in-the-wild setting did not allow as many identifications, nor as many rounds, we used the number of director descriptors rather than the number correct as the entrainment measure. In Example 2, descriptors would include *blob*, *concrete thing*, *sculpture*, and *jellyfish*. We chose the number of director descriptors instead of other measures (e.g., number of matcher descriptors) because directors' descriptors were less subject to cell phone reception problems and because director and matcher descriptors were correlated and therefore similar (see Coding section). A higher number of descriptors indicates greater difficulty communicating about the object. As the number of descriptors shorten, communication becomes more efficient.

If entrainment occurs in this naturalistic setting, round 2 should have fewer director descriptors for any target object than in round 1. That is, directors' descriptions in round 2 will be shorter if and only if entrainment has occurred in round 1. This is an oft-reported finding in the literature, which is demonstrated in laboratories by showing that descriptions shorten over rounds in dialogues (with entrainment) but not monologues (no entrainment; e.g., Clark & Wilkes-Gibbs, 1986). The important predictor variable is the *round*, not the primes (what one person says) and targets (what the other person says). Research on entrainment has used both rounds and primes within a dialogue. We use rounds here.

The measure of entrainment – fewer descriptors in round 2 than in round 1 – accommodates implicit agreement across interlocutors. Most of the time, deciding the label for an object (e.g., a *blob*) is not accomplished with explicit statements (e.g., "Let's call this a blob") nor explicit confirmations (e.g., "Do you agree?" "Yes, I do"). It is accomplished by reducing the number of descriptors used (e.g., *blob*, *concrete thing*, *sculpture*, and *jellyfish* in round 1 and *blob* in round 2, as in Example 2). In past work, measurement of verbal agreement across interlocutors (such as displayed by the reduction of descriptions across rounds) has not been a typical assessment of entrainment; the typical assessment is accuracy of object identification, such as how many objects were correctly picked out of an array. One reason verbal agreement has not been used is that it requires a high amount of judgement calls from coders. To illustrate, when one interlocutor uses the phrase "like a t," while the other interlocutor uses the phrase "t-shaped," should these be counted as the same descriptor, or different ones? When one interlocutor uses the phrase "copper thing" and the other uses each word separately ("thing" at one point and "copper" at another point), is that two examples of entrainment or one? Accuracy in this kind of assessment relies on human judgements: A computer may be able make these comparisons consistently, but its matches would be extremely sparse. In the Methodology section we provide analyses of how often directors' and matchers' word choices overlapped. Then, to measure entrainment, we used the number of director descriptors as our main dependent measure.

In addition to assessing whether director descriptors shortened across rounds, we examined whether participants' acquaintanceship status (they were either friends for at least one academic quarter or complete strangers) and extraversion influenced communicative efficiency. Because the successful completion of referential communication tasks is thought to be reliant on the creation of conceptual pacts and shared common ground (Brennan & Clark, 1996), we predicted that friends would be more efficient than strangers (analysis 1). Additionally, we tested how communicators behaved across multiple rounds of the task (rounds 1 and 2). Round 1 relied on directors' and matchers' abilities to negotiate a referring expression for a novel target with each other, a process which was made more challenging for matchers walking around downtown without a map. Round 2 relied on directors using descriptions of previously identified targets in ways that matchers would recognize, a process that was dependent on the participants' remembering the negotiated label. We expected round 1 performance to be more likely to involve interpersonal and personality factors (analysis 2) and round 2 performance to be more influenced by conversational history, such as how quickly the pair found the targets the first time (analysis 3).

2 Methodology

In this section, we provide details on the methods we used for this study.

2.1 Participants

Forty-eight pairs of UC Santa Cruz participants' audio recordings were analyzed in this study (24 friend pairs and 24 stranger pairs). Those recruited to be in friend pairs were asked to bring a friend, which was defined as someone they spoke to regularly and had known for at least one academic quarter (10 weeks). An additional 26 pairs' recordings were not assessed because they failed to find the minimum number of targets (i.e., at least 8 of 10) needed to ensure a similar number of trials across pairs (19 pairs) or because they experienced experimenter, participant, or technical errors such as equipment failing to record, participants' taking photos in the wrong order, or experiencing particularly poor cellular connection quality during the task, which included cases where the phone call was dropped more than once (7 pairs). Participants received either course credit or a $10 gift card after participation. Participants were screened on native language so only native English speakers were included in this study. Table 1 shows the gender composition of the 48 analyzed pairs, with females outnumbering males (58% to 42% respectively).

Table 1. Gender composition of analyzed pairs ($N = 48$).

	Female Director	**Male Director**
Female Matcher	21 (43.8%)	7 (14.6%)
Male Matcher	8 (16.7%)	12 (25%)

2.2 Materials

Downtown Santa Cruz, California, features many public art installation pieces such as sculptures, murals, and mosaics. About 40 pieces of abstract and non-abstract art were initially identified as potential targets. Ten research assistants were asked to describe each object in detail and the number of descriptive words used (*descriptors*) was tallied and averaged. The objects were ranked by number of descriptors used; the ten with the greatest variety of descriptors were chosen as targets, to ensure that multiple conversational turns would be required to identify the objects. All ten targets were located within a two-by-six block area. Most targets were close to other pieces of art that often needed to be explicitly eliminated by the matcher as a potential target; this sometimes included finding specific, unique sections of murals and multi-panel mosaics. During the yearlong data collection process, three pieces were removed by the city. We replaced those targets with new pieces that were geographically close to the originals.

The 10 selected targets were split into two lists (A or B) with five targets each, with an attempt to balance the maximum distance participants would be required to walk in order to take photographs of all five targets.

2.3 Procedure

Director/matcher pairs worked together to find art targets in downtown Santa Cruz. Directors were located in a campus lab and matchers were located downtown. There were two rounds of art identification. In each round, the director received five targets in sequence, describing each one to

the matcher. When the target was found, the matcher took a photo of it and the director hit a key to advance to the next art object. The experimenters in the two locations both had their own cell phones which they could use to indicate issues with the study, such as a late participant or study cancellation due to rain.

Directors met an experimenter in a campus lab while matchers met an experimenter at a café a block away from the center of the downtown area where all targets were located. Before establishing a call between participants, experimenters gave a short orientation for the devices the participants would be using. After the call was established, the experimenters left the participants alone to do the practice trial and experiment (the campus experimenter waited in another room while the downtown experimenter stayed behind at the café). The practice trial used a non-abstract statue as a target near the center of the downtown area, allowing participants to become acclimated to the equipment and task setup while also placing them close to the trial targets.

Directors were connected to matchers via voice-only Skype in an enclosed computer booth. Directors used the Skype application on a computer, but matchers received the call as if it were a typical cell carrier-based voice call. Matchers carried a smartphone and a separate digital camera. Both participants were given the option of wearing headsets, but most participants opted against them (matchers spoke holding the phone to their ears, directors were on speaker phone). The conversation was recorded using Audio Hijack Pro 2, an in-line audio recording program.

During an experiment trial, directors were shown a photo of a single target object alongside a map with the target's location highlighted in green (see Figure 1). Photos of the target stimuli and a map were presented to the director using SuperLab 4.5 (Haxby, Parasuraman, Lalonde, & Abboud, 1993). The images were displayed on a single computer screen. The map of downtown Santa Cruz used Google Maps' satellite imagery, but with indicators manually inserted to show the locations of targets. Matchers did not see photos of the stimuli. Matchers were also not given maps. Matchers' self-reported familiarity with downtown Santa Cruz was not significantly correlated with the number of descriptors or turns they or their partner used, suggesting that they were indeed reliant on the director to get them to where they had to be and/or had little preexisting knowledge of specific art installations.

Each new target was accompanied with the same map but with new locations highlighted and old locations in grey. Target order was randomized, and each target was given an 8-minute time limit. The goal of the time limit was to prevent the entire hour from being spent on searching for one piece of art. (We note here that five objects found twice with a maximum of 8 minutes each equals 80 minutes, which was 20 minutes beyond the hour allowed for the activity.) If directors failed to hit the key to advance to the next target within 8 minutes, the experiment automatically moved on to the next target and participants had to stop their search for the timed-out target. This was done in order to motivate participants to progress as much as possible through the ten trials, rather than getting stuck on one. After participants searched for all five targets once (round 1), targets were re-randomized and presented again for the pairs to find a second time (round 2). Participants were instructed to not deviate from the order set for them; the handful of pairs that did were excluded from analyses. Matchers were instructed to take pictures of the targets as they located them, which were then checked by the experimenters after the experimental session was concluded. Although directors had no access to the photos, directors and matchers confirmed verbally that they had found the right targets. The fact that directors had no access to the photos conceptually mirrors both the Map Task and the Tangram Task, neither of which typically confirms for participants that they have arrived at the same understanding with evidence beyond verbal confirmation (for example, directors do not typically see the tangrams matchers select). Generally, matchers were at ceiling in finding targets.

When they had gone through both rounds, directors led matchers back to the café, after which the call was disconnected. Both participants were separately given post-experiment questionnaires which included a single item asking them to rate their familiarity with the downtown area on a 7-point Likert-type scale, five items on their experience doing the study, as well as the Ten Item

Personality Inventory (TIPI; Gosling, Rentfrow, & Swann, 2003) which was used to assess extraversion. Participants were asked to rate themselves on pairs of personality-related words on a scale of 1 (*Strongly Disagree*) to 7 (*Strongly Agree)*.

Figure 1. An example of a director's screen during the task. The map is non-interactive and has grey indicators for potential target locations. The green indicator shows the location for the current target and the red indicator shows the experiment starting point.

2.4 Coding

We begin this section by describing the data included in analyses, and then how the data was coded to measure communicative efficiency. The nature of the planned analysis required that only the data for pairs who had found at least 8 of the 10 targets would be analyzed for this study, at least four targets in each of two rounds. Inability to find a target often had a cascading negative effect on performance: Participants generally either became flustered when the experimental procedure moved on after eight minutes spent on a target, or they ignored the time limit and kept looking for the target, which then threw off timing of the rest of the trials. Table 2 shows the number of friend pairs and stranger pairs who found 8, 9, and 10 targets.

Table 2. Number of targets found by pairs.

Number of Targets	Friend Pairs ($N = 24$)	Stranger Pairs ($N = 24$)
8 targets	2	7
9 targets	8	4
10 targets	14	13

In the Participants section we noted that participants who had problems with their photos were excluded. The problem with including pairs who took photos in the wrong order is that there was no way to retroactively determine whether they found the correct objects on each trial or if they had found targets of prior trials during later trials. There were also problems with participants who took overly wide shots because these shots included multiple potential targets.

There were two dependent measures available to quantify communicative efficiency, the number of directors' descriptors per target and the number of matchers' descriptors per target. These two variables were moderately to strongly correlated with each other (round 1 $r = .23$; round 2, $r = .58$), suggesting that they would operate similarly for our analyses. We chose directors' descriptors because they were less sensitive to the pitfalls of cellular communication on a street; for example, there were several times that matchers' reception cut out on the recordings.

Research assistants annotated director descriptors, matcher descriptors, and counted the number of turns taken for each piece of target artwork found. Descriptors were defined as unique-within-round adjectives or nouns that were descriptive of the target objective, such as colors, shapes, media (e.g., painting, sculpture, metal, stone) and patterns (e.g., striped). Coders were told that descriptors should generally be single or hyphenated words, such as "purple" and "egg-shaped." Despite this, coders found that many descriptors required longer phrases. Adjectives that referred to a portion of the artwork needed to retain their noun, e.g., "green background," "multiple curves," and "pointy nose." Non-hyphenated compound nouns needed to be included, e.g., "stick figures," "triangle shape." Other phrases were used to disambiguate between attributes of the work itself and its position within the greater context, e.g., "on top of a platform," "on a cinderblock," "part of a bigger mural." Even longer phrases simply lost meaning when they were artificially shortened, e.g., "if it had hands, they would be doing spirit fingers." Allowing research assistants to be more dynamic with their coding made sense: Even in the tangram task, descriptors are longer in the first trial (e.g., "a person who's ice skating" vs. "the ice skater"). To avoid arbitrarily determining what counted as synonyms, similar descriptors were counted separately. For example, the uses of "turquoise," "blue," and "kind of blue" were treated as unique to allow them to be considered as terms that could be entrained upon despite a possible lack of reuse of the term by a matcher (as when a matcher responds to a series of descriptors by accepting but not repeating them, such as by saying "ok"). Our goal was to capture the widest range of descriptors without over-reliance on judgement calls (e.g., Is *turquoise* the same as *blue*?). Each unique descriptor was only counted once per trial. This is due to the constraints of running a task "in the wild": Noisy streets and occasionally poor cellular connection meant that there was a lot of repetition of single words or short phrases by both director and matcher. Counting repeated words as separate instances of entrainment would conflate entraining on descriptors with ensuring correct audio transmission, adding unwanted noise to the analysis.

To determine whether this coding scheme was similar across coders, two research assistants coded round 1 director descriptors for 12 of the 48 participant pairs. Because descriptor coding was an open-response task, we chose three measures to understand how similarly descriptors were coded. First, we used a Pearson correlation to determine whether the number of descriptors coded by one research assistant for one trial would predict the number of descriptors coded by the other research assistant for the same trial. A positive correlation would suggest that when one coder recorded more descriptors for a trial, the other coder would in general also record more descriptors. A count-based analysis of similarity between descriptors is especially useful in this context, as descriptor count is also our measure of entrainment. The number of descriptors that each research assistant coded was highly correlated with the other research assistant's coding, $r(58) = .876$, $p < .001$.

Second, we performed a flexible human-based procedure where a third research assistant compared the two coders' work using a strict-matching method (similar to how a Python script might compare strings, but in a slightly more forgiving way) and a flexible-matching method that asked for just the meaning to be similar enough (even more forgiving). Each of these match coding

schemes is described in detail below. The match coding instructions that were given to research assistants are included in Appendix A.

The strict method of coding was similar for single-word and multiple-word descriptors. In the strict method of coding single word descriptors, the third research assistant determined whether each descriptor was reported by both coders. When close or partial matches occurred, they determined whether the meaning was similar enough to be treated as a match based on whether the root of the word was the same, such as "spikes" and "spiky" or "yellow" and "yellowish." For multiple-word descriptors, they determined whether content words were identical. The third research assistant followed the example that "surrounded by plants" and "plants are surrounding it" have two identical content words, *plants* and *surround*. We also specified that negation within one of two multiple-word descriptors (e.g., "holding hands" and "not holding hands") should not be counted as matching. They compared single-word descriptors to multiple-word descriptors in the same way that they compared multiple-word descriptors. The third research assistant's analysis found 38.98% agreement between the two original coders, averaged across all 12 analyzed participant pairs, for the strict method of coding.

Strict descriptor matching is likely to under-represent the actual levels of entrainment due to its rigid rule set. For example, the strict rule set would not allow "rocks" and "stones" to match. For a deeper look at inter-coder overlap, the third research assistant made more thoughtful judgment calls on whether a descriptor's meaning was similar enough to be considered the same. Using these rules, the third research assistant's analysis found 49.04% agreement averaged across items.

Finally, in order to achieve a more objective measure of the similarity between the research assistants' descriptor coding while also factoring in semantic content, we relied on recent work showing that word embeddings provide state of the art results on Semantic Textual Similarity tasks, such as our task of comparing the lists of phrases in our descriptor pairs (Liu, Ott, Goyal, Du, Joshi, Chen, Levy, Lewis, Zettlemoyer, & Stoyanov, 2019). Word embeddings capture semantic similarities between words, generalizing beyond the particular words used in the descriptors (Reimers & Gurevych, 2019; Li et al 2020)). Using Roberta LARGE embeddings[1], we measured the cosine similarity of the document embeddings of the lists of director descriptors coded by the research assistants. We first directly compared descriptors from the same conversation for each Artwalk object as illustrated by Table 3. Over all 60 samples of "same Artwalk object, same conversation" we get a high average similarity (M = .80, SD =10). We then compared these measures of cosine similarity with a random selection of 60 descriptions given by one research assistant matched to a description given by the other research assistant for the same Artwalk object, but from a different conversation (different dyads). This is illustrated by the pairs of descriptor lists in Table 4. The cosine similarities for "same Artwalk object, different conversation" were significantly lower (M = .44, SD = .15, $t(59) = 15.65$, $p < .001$). This difference clearly shows that research assistant coders coded more similarly when coding the same conversation than coding different conversations about the same Artwalk object. In order to establish a random baseline for cosine similarity for Artwalk objects, we also created a random selection of 60 descriptions from one research assistant and 60 descriptions from the second research assistant, for different Artwalk objects in different conversations. See Table 5 for examples. Here the cosine similarities for "different Artwalk object, different conversation" were again significantly different (M = .27, SD = .12, $t(59) = 26.65$, $p < .001$).

While it is challenging for humans to agree on descriptors, embeddings show that the descriptors they report are similar at levels far above chance. We also note that our main research question (whether entrainment occurs) will be measured by the *number* of descriptors coded, rather than what those descriptors actually are.

[1] https://github.com/MartinoMensio/spacy-sentence-bert/

Table 3. Semantic textual similarity scores for the same object within dyads.

Object	Coder 1	Coder 2	Similarity
Emo Penguin	emo penguin, spray painted, beanie, white, sad	emo, penguin, spray painted, outline, beanie, white,	0.922
Trumpet Bench	across from parking garage, middle of the block, mural, wooden bench, dog legs, instrument, trumpet, red and blue ribbons, painting on cement, animal legs, carved legs, tree, white low fence	mural, painting, bigger, wooden, bench, two, legs, animal-like, dog legs, instrument, trumpet, red, blue, ribbons, cement, carved, behind, white, low, fence	0.862

Table 4. Semantic textual similarity scores for the same object across different dyads.

Object	Coder 1	Coder 2	Similarity
Kimono	moldy, t shape, red, big, green, towards borders, in front of that Indian place	weird statue, torso, across the street from Om store, metro center side	0.339
Vader Helmet	black, shiny, marblish, circular on top, flat bottom, carved stripes, in front of clothes store	dome-shaped, sculptury, concrete, black, seal, smooth	0.299

A phenomenon we did not analyze was direction-giving ("on Lincoln Street," "turn left and walk one block"). Traditional lab-based referential tasks do not generally include route-finding, and we wanted to draw a more direct comparison between the Tangram Task and the Artwalk task. Other researchers have also opted to focus on specific target descriptions when analyzing communication produced in a real-world wayfinding and target-identification task (e.g., Brennan et al., 2013).

Number of turns, which was only used in analysis 3, was the combined number of turns taken by both director and matcher from the time the director started any conversation about the next target to the time when the matcher indicated that they had found the target. Because some turns were partially about directions and partially about descriptions, instead of determining how much of a turn needed to be dedicated to description to count, we included partial turns and turns dedicated to direction-giving in this variable.

Table 5. Semantic textual similarity scores for different objects and across different dyads.

Object	Coder 1	Coder 2	Similarity
Coder 1 object: Kimono Coder 2 object: Pi Signs	rock, moose, green, brown	right side of street, blue background, rectangle, two parallel lines, white, bars, connect to other side, in between, strip	0.119
Coder 1 object: Mosaic Coder 2 object: Emo Penguin	in a parking lot, to the right, collage, tile, blue, green, orange, red	penguin, little, boy, white, beanie, pointy, black, shading, helmet, large, buildings, facing, left	0.107

To summarize, the number of director descriptors was used in analyses 1 and 2, and the number of director descriptors and turns was used in analysis 3.

2.5 Examples of Entrainment and Descriptor Overlap

In this section, we illustrate what is meant by entrainment in dialogue using examples from the corpus. We also describe how much descriptors overlapped between directors and matchers, followed by a discussion of the relationship between entrainment and overlap.

In Figure 2, the matcher uses the same words as the director, "a bird with a beanie," as part of the process of grounding on descriptors that identify the art.

1 M: is it on the left the uh little mural?
2 D: yeah there is a picture of a bird with a beanie on
3 M: a bird with a beanie on you said?
4 D: yeah it is like a mopey gray sad bird
5 M: oh okay yeah I see it
6 D: yeah
7 M: yeah it is on the wall
8 D: okay cool

Figure 2. An example of a matcher copying a director in round 1.

The Figure 2 example was from round 1. In round 2, the director shortened the description to "the bird" saying, "uh so you're gonna go back to the bird one um so walk up Pacific," which was accepted by the matcher with "alright I just got a picture of it." This shows that by round 2, the director and matcher had entrained on "bird" for this art piece, and that the matcher did not need to repeat the word "bird" for entrainment to occur.

The descriptor "a beanie" was also used by a different director-matcher pair, as seen in Figure 3 lines 17 and 24. But unlike Figure 2, the director and matcher in Figure 3 made a lot of other conversational contributions that illustrate the breadth of information sifted through in this corpus to quantify entrainment, including directional information (e.g., line 12), spatial information (e.g., line 7), and conversational coordination (e.g., line 2, lines 31-32).

In the second round of the Figure 3 pair, the target was described exclusively with the director's words, entraining on "sad penguin thing with the beanie," as shown in Figure 4, but without the matcher's repeating any of the words in "sad penguin thing with the beanie."

1 D: you're looking for like a emo penguin
2 M: sorry what?
3 D: it's like a emo penguin or something I'm not really sure exactly what it is
4 M: okay [chuckle]
5 D: yeah
6 [silence 65s]
7 M: how far down did you say? I'm on by Church right now
8 D: by by Church Street?
9 M: yeah
10 D: it's it's the next street
11 M: okay
12 D: yeah [silence 3s] uh when you get to Locust you wanna cross the street um cross Locust or a take a right on Locust and then cross it
13 M: okay so go the next block but on Locust
14 D: yeah
15 M: okay
16 [silence 7s]
17 D: and um you're so what you're looking for it's like a it looks like it was spray painted um it's like an outline I think it's like a penguin wearing like beanie and it's colored in white mostly
18 M: okay um
19 D: and it should be um it looks like it's probably there should be two buildings on Locust Street from what I can see on that side of the street so it's not the one it's not the one on the Cedar side it's the one on the Pacific side and it's
20 M: alright
21 D: on the corner of the building that's closest to the middle of the block
22 M: okay [silence 4s] so you said look like oh okay I think I see it
23 D: you see it?
24 M: wearing a beanie?
25 D: yeah
26 M: okay and okay so do you just want that one specific penguin looking like thing or do you want all of them?
27 D: uh it only has the one penguin it's only it's like cropped to where it's only the one white penguin on mine so
28 M: okay so
29 D: so looks pretty sad
30 M: cool alright
31 D: got it?
32 M: yeah

Figure 3. An example of object entrainment surrounded by directional and conversational coordination in round 1. In this exchange, one of our coders generated the following list of descriptors for the director: *emo*, *penguin*, *spray painted*, *outline*, *beanie*, and *white*. Note that *penguin* only occurs once in this list, as we asked coders to only record unique descriptors (only ones that hadn't already occurred in the trial).

1 D: got that one? so now you gotta go back to the sad penguin which is um on Locust Street so I mean you can go to Pacific or Cedar and just head head that way
2 M: okay
3 D: away from Pergolesi
4 M: alright wait which one am I getting now?
5 D: um the sad penguin thing with the beanie
6 M: oh okay okay okay yeah

Figure 4. Round 2 of the Figure 3 example showing drastic reduction in words used to identify the art.

Both friends and strangers entrained across rounds 1 and 2. Figure 5 illustrates the same target described during the first and second rounds of a friend pair and a stranger pair (asterisks indicate overlap). Examples were chosen for brevity and clarity in round 1. They came from pairs where the matcher became reasonably confident in the identification of the target quickly.

Friends describing the spiky rock target in the first round

1 D: you'll be pretty much walking pretty much towards the end of Cooper Street and what you're looking for, um pretty much they're spiky rocks, like, they're rocks with little spikes on them
2 M: it's a what?
3 D: they're rocks, like, kinda like oval-shaped rocks with spikes on them, they're, you you'll, you'll know what I mean
4 M: huh? no, you were saying it's an oval-shaped rock
5 D: with spikes on them [yawn]
6 M: wait, it's what?
7 D: spikes, like, sp- like thorn spikes, like spiky stuff like s-
8 M: is it red?
9 D: one of the rocks is charcoal gray, one of the rocks is like a light brown, the other rock is *like*
10 M: *oh do* the- do they do they have have a little like slots on them?
11 D: yeah, like little spikes on them yeah
12 M: okay, I found them

Same friends describing the spiky rock target in the second round

14 D: okay, um last one, go back to Cooper Street for the spiky rocks
15 M: hey hey to the- or- to the rocks again?
16 D: yeah, the ones that we yeah.
17 M: okay, to the rocks and then it's the last one, that one mural
18 D: yeah, I think that'd be the last one
19 [matcher and director joke and matcher says something unintelligible to a third party]
20 M: alright, I got it

Strangers describing the spiky rock target in the first round

21 D: okay, so you head down Cooper
22 M: *uh huh*
23 D: *and* uh it's gonna be on the left side er- it's gon- once you- at the end of Cooper if *ya*
24 M: *uh huh*
25 D: on, the left of you, the left side, it's at the corner, there should be like these three uh it looks like these three rocks one's- the the- it goes from like one's small, one's medium, and one's one's large
26 M: oh yeah yeah, with uh blue spikes and *yellow spikes*?
27 D: *yeah yeah yeah* those are the ones
28 M: perfect
29 D: okay

Same strangers describing the spiky rock target in the second round

30 M: okay, I got that picture, where to next?
31 D: next one is the uh three wa- the three stones the blue one with the spikes
32 M: are we just revisiting all of them?
33 D: yeah, we're *revisiting all of them*
34 M: *[laughing]*
35 D: so it's on Cooper Street, you're right next to it
36 M: alright, I was like wait a minute, *these objects*
37 D: *[laughing]*
38 M: look familiar
39 D: yeah
40 M: okay, well at least that one's close
41 [task-irrelevant conversation as matcher takes photo]
42 M: anyway okay, I got that last picture, so now what?

Figure 5. A pair of friends and a pair of strangers describing the same art across rounds 1 and 2.

All four of these participants used similar concepts to describe this public art, (1) rocks or stones (e.g., lines 1, 14, 25, and 31) and (2) having spikes (e.g., lines 1, 14, 26 and 31).

A close examination of directors' and matchers' descriptors was conducted on half the data to assess the degree to which the descriptors overlapped. Unsurprisingly given that they had the information about the objects to be identified, directors produced more descriptors than matchers: Directors' descriptors account for 64.63% of the total descriptors. We also asked two coders to assess how often directors' and matchers' descriptors overlapped using the two sets of rules that were used above to measure similarity, both using the strict and the flexible approaches. Coders' ratings of the number of director descriptors matched to matcher descriptors across 60 trials were highly correlated for both the strict set of rules, $r(58) = .94$, $p < .001$, and the flexible set of rules, $r(58) = .79$, $p < .001$.

In the strict method of coding single word descriptors, both raters determined whether a director's descriptor was also used by the matcher, such as "rocks" in Figure 5. As before, when close or partial matches occurred, coders determined whether the meaning was similar enough to be treated as a match, based on whether the root of the word was the same. Also as before, for multiple-word descriptors, coders determined whether content words were identical including that negation should not be counted as matching. Coders compared single-word descriptors to multiple-word descriptors in the same way that they compared multiple-word descriptors. Strict coding resulted in 12.17% overlap averaged across coders and items.

As before, strict descriptor matching is likely to under-represent the actual levels of entrainment due to its rigid rule set. For a deeper look at entrainment, coders made more thoughtful judgment calls on whether a descriptor's meaning was similar enough to be considered entrainment. Flexible coding resulted in 21.99% overlap average across coders and items.

The strict and flexible overlap analyses show that people do in fact overlap in the words chosen to describe the target art, which illustrates one form of entrainment. Entrainment does not require overlap in words, however. Entrainment is also visible in Figures 2 through 5 through the use of backchannels acknowledging agreement on a description. In Figure 4 line 6, the matcher responds "oh okay okay okay yeah" to the descriptor "sad penguin" without repeating any descriptor words. In Figure 5 lines 30 through 42, the matcher doesn't even confirm that the "three stones" were found — the confirmation is implicit in the matcher's clarification that the objects were being identified a second time coupled with the matcher's final remark, "Anyway okay, I got that last picture, so now what?" The entrainment occurs through the taking of the picture, not with the production of words like *rock*, *stone*, or *spiky*.

3 Results

We tested whether acquaintanceship status and extraversion influenced how quickly (defined by number of descriptors) pairs would entrain (analysis 1) and examined how efficiency differed in round 1 (analysis 2) and in round 2 (analysis 3). Because there was no evidence of a difference in performance between list A ($N = 23$) and list B ($N = 25$; $t(46) = -0.65$, $p = .52$, 95% CI [-2.90, 1.49]), lists were collapsed in these analyses. Descriptive statistics broken down by acquaintanceship can be found in Table 6. We note here that that there was no evidence of a difference in director descriptor use before and after the replacement of two of three targets during our data collection period (due to the city changing its public art displays), $t(46) = -0.73$, $p = .47$, 95% CI [-3.05, 1.43] (because the third target was switched out very late during our data collection with only a few participant pairs discussing it, we did not break the data down to test whether the third object made a difference).

Overall, there was a slight negative skew (-0.21) in the director extraversion data, with more participants rating themselves as being highly extraverted than highly introverted. The scores were non-normally distributed (Shapiro-Wilk = 0.95, $p = .03$) but the distribution was still roughly bell-shaped.

Table 6. Basic descriptive statistics on variables of interest broken down by acquaintanceship status.

Variable		Friend Pairs (N = 24)	Stranger Pairs (N = 24)
Round 1 Director Descriptors	Mean	12.53	11.93
	SD	3.90	3.66
	Median	12.70	12.00
Round 1 Turns	Mean	20.01	18.67
	SD	5.49	8.68
	Median	19.68	16.93
Round 2 Director Descriptors	Mean	4.34	3.99
	SD	1.91	1.64
	Median	4.20	3.68
Round 2 Turns	Mean	8.32	7.05
	SD	3.59	2.75
	Median	7.00	7.00
Director Extraversion	Mean	4.48	4.81
	SD	1.42	1.47
	Median	4.00	5.00

3.1 Analysis 1

With analysis 1 we tested whether extraversion and acquaintanceship predicted the number of director descriptors used overall, as well as whether extraversion and acquaintanceship had differential effects in round 1 and round 2. Preliminary analyses indicated no evidence of a difference between director extraversion between friends (M = 4.48, SD = 1.42) and strangers (M = 4.81, SD = 1.47) in our sample, $t(46) = -0.80$, $p = .43$, 95% CI [-1.17, 0.50]. This suggests that friend pairs who participated did not inherently differ from stranger pairs with respect to sociability or outgoingness.

To test whether acquaintanceship status and extraversion influenced how quickly (defined by number of descriptors) pairs entrained, a hierarchical linear regression was performed using director descriptors as the dependent variable, round and acquaintanceship as binary categorical predictors, and director extraversion as a continuous predictor (centered; Cohen, Cohen, West, & Aiken, 2003). Two-way interaction terms between the three predictor variables were also entered. This model (Model A) predicted about 69.62% of the variance in director descriptors, $F(6,95) = 37.29$, $p < .0001$, adj-$R^2 = .70$, though it should be noted that the interaction terms introduced a fairly high level of collinearity. Acquaintanceship ($b = -1.42$, $p = .43$) and director extraversion ($b = 1.10$, $p = .10$) were not significant predictors of the number of director descriptors. However, round was a significant predictor ($b = -8.31.10$, $p < .001$), indicating that directors did use shorter descriptions in round 2 than in round 1. This result replicates, in the wild, the oft-demonstrated laboratory finding that referring expressions are shortened during entrainment in referential communication tasks.

There were two interactions. One interaction was between acquaintanceship and director extraversion ($b = 1.14$, $p = .005$). Whereas directors in friend pairs used the same number of descriptors regardless of whether they were introverted or extraverted, the number of director

descriptors for stranger pairs was sensitive to director extraversion. Specifically, stranger directors who scored higher in extraversion used more descriptors than those who scored lower in extraversion. The second interaction was between round and director extraversion (b = -0.85, p = .04). director extraversion only made a difference to the number of descriptors produced in round 1 but not in round 2. This possibly suggests that round 1 and round 2 are fundamentally different when it comes to the influence of extraversion on performance. There was no interaction between round and acquaintanceship (b = 0.53, p = .64). See Table 7.

Table 7. Regression table for Analysis 1: All Director Descriptors - Model A. **Sig. at .01 level; ***Sig. at .001 level.

	B	**SE**
Round	-8.31***	0.78
Acquaintance	-1.42	1.79
Director (Dir) Extraversion	1.11	0.66
Acquaintance * Dir Extraversion	1.14**	0.40
Acquaintance * Round	0.53	1.13
Round * Dir Extraversion	-0.85	0.40
Adjusted R^2	0.70	
F for Model	37.29***	

3.1.1 Analysis 1 Discussion

In analysis 1, we found that neither extraversion nor acquaintanceship alone influenced the number of descriptors that the director used, though round did influence it. Directors used fewer descriptors in round 2 than in round 1, which replicates the findings of lab-based tangram tasks. We also found that extraversion influenced directors from stranger pairs but not friend pairs. Directors from stranger pairs who were more extraverted used more descriptors than the less extraverted directors. But there was no evidence that friend directors used a different number of descriptors depending on whether they were high or low on extraversion. This was unexpected because extraversion is generally associated with greater talkativeness (Dewaele & Furnham, 1999) and a greater volume of novel ideas (Jung et al., 2012).

We also looked at whether extraversion and acquaintanceship interacted with round. This was used to test whether round 1 and round 2 should be analyzed separately. Because round interacted with at least one of the predictors (director extraversion), we examined the two rounds using separate regression models for the remainder of the analyses of this study. In analysis 2, we examined round 1 using director extraversion and acquaintanceship as predictors. In analysis 3, we explored the influence of extraversion and acquaintanceship on the number of descriptors used by

directors in round 2, as well as two predictors stemming from round 1: number of director descriptors in round 1, as well as the number of turns in round 1.

3.2 Analysis 2

With analysis 2, we tested the influence of director extraversion and acquaintanceship on the number of director descriptors produced in round 1. The predictors used were director extraversion (b = 0.07, p = .89), acquaintanceship (b = -0.89, p = .39), and their interaction (b = 1.50, p = .04). This model, Model B, accounted for about 14% of the variance in director descriptors, $F(3,44)$ = 3.46, p = .02, adj-R^2 = 0.14. See Table 8.

Table 8. Regression table for Analysis 2: Round 1 Director Descriptors - Model B. *Sig at .05 level.

	B	SE
Acquaintance	-0.89	1.02
Director (Dir) Extraversion	0.07	0.51
Acquaintance * Dir Extraversion	1.50*	0.71
Adjusted R^2	.14	
F for Model	3.46*	

We used simple slopes analysis (Cohen et al., 2003; Preacher, Curran, & Bauer, 2003) to examine whether acquaintanceship moderated the extent to which extraversion influenced the number of director descriptors used in round 1. Simple slopes analysis probes an interaction by examining the regression of the dependent variable (director descriptors) on a predictor variable (extraversion) at different values of a moderator variable (acquaintanceship), which in this case is dichotomous. The significance test assesses whether the slope of the regression line for either friend or stranger pairs differs from zero: that is to say, whether a decrease or increase in director descriptors is associated with higher or lower director extraversion.

In order to graph the interaction, high extraversion was calculated using one standard deviation above the mean of our sample while low extraversion was calculated as one standard deviation below. This was both parsimonious and intuitive, given that there was no evidence that director extraversion differed between the two acquaintanceship conditions and given that our sample's extraversion is in line with typical Ten Item Personality Inventory extraversion scores (Gosling et al., 2003). One SD above the mean was a score of 6.08 and one SD below the mean was 3.21 (on a scale of 1-7, where the midpoint is 4).

There was no evidence that friend pair directors differed in the number of director descriptors produced based on how introverted or extraverted the director was, but there was evidence that stranger pair directors differed. The less extraverted the stranger director, the fewer descriptors they used. The stranger slope deviated from zero, stranger slope = 1.57, $t(46)$ = 3.17, p = .003, 99% CI [0.24, 2.90], but the friend slope did not, friend slope = 0.07, $t(46)$ = 0.13, p = .89. See Figure 6.

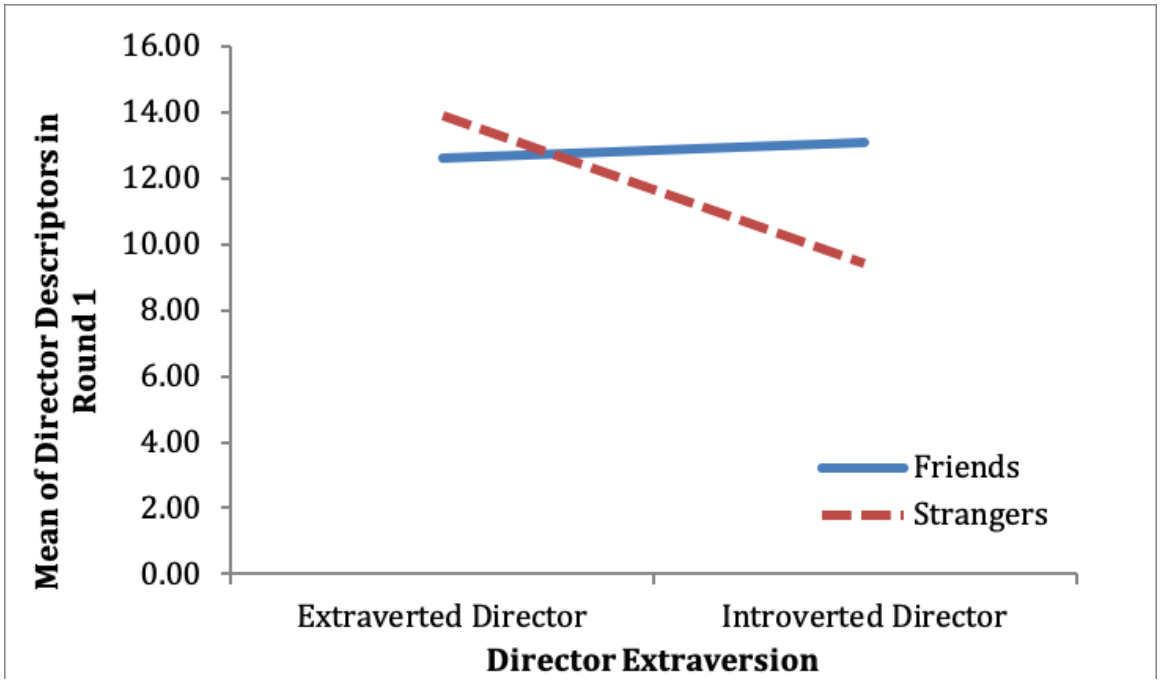

Figure 6. Simple slopes graph for friend and stranger pairs for round 1 director descriptors on extraversion.

3.2.1 Analysis 2 Discussion

In analysis 2, we found that there was no direct effect of being previously acquainted with a partner on performance. We also found no simple relationship between director extraversion and the number of descriptors used. Instead, we found that acquaintanceship moderated the influence of extraversion, such that extraversion only had an effect when participants were strangers. There was no evidence that friend directors differed in the number of descriptors produced depending on how introverted or extraverted they were, but there was evidence that stranger directors did. The less extraverted the stranger director, the fewer descriptors they used.

Though we found that an interaction between acquaintanceship and extraversion accounted for some of the variance in descriptors in round 1, we predicted that this pattern would not hold for round 2. Once participants believed they had established a conceptual pact for specific referents, the task became a more straightforward test of whether the matcher was able to match a referring expression to one of the targets previously found in round 1. We predicted that round 2 performance would be related to how participants did in round 1: The less confusing and more straightforward round 1 performance, the fewer number of descriptors the director would need to use in round 2.

3.3 Analysis 3

With analysis 3, we explored the influence of director extraversion, acquaintanceship, the number of director descriptors in round 1, and the number of turns in round 1 on the number of director descriptors produced in round 2. This was an exploratory analysis because we did not have specific hypotheses. We wanted to explore whether the speed at which people were able to identify art – fewer turns in round 1 – influenced the effort required to identify art in round 2 – fewer descriptors in round 2.

We ran the same model that we used for analysis 2 (which was Model B in analysis 3: director extraversion, acquaintanceship, and director extraversion * acquaintanceship interaction) on the number of director descriptors used in round 2. In Model C we see no main effects or interactions across these variables. Model D went a step further and added round 1 director descriptors, as it

was reasonable to assume that the number of descriptors used in round 1 would influence the number of descriptors used in round 2. Model D also added two two-way interactions: Round 1 director descriptors * director extraversion and round 1 director descriptors * acquaintanceship. All predictors were centered, except for the binary acquaintanceship predictor. These additions resulted in a regression model that did not reach significance, $F(6,47) = 1.84$, $p = .12$. As a result of the results of Model D, both director extraversion and round 1 director descriptors were dropped from the model. Model E substituted round 1 turns in place of round 1 director descriptors as the measure of efficient communication in round 1. Round 1 turns was logarithmically transformed, as it was moderately skewed. Model E, which includes acquaintanceship, round 1 turns (log), and an interaction between acquaintanceship and round 1 turns, is able to account for about 10.4% of the variance in round 2 director descriptors, $F(3,44) = 2.83$, $p = .05$, adj-$R^2 = 0.10$.

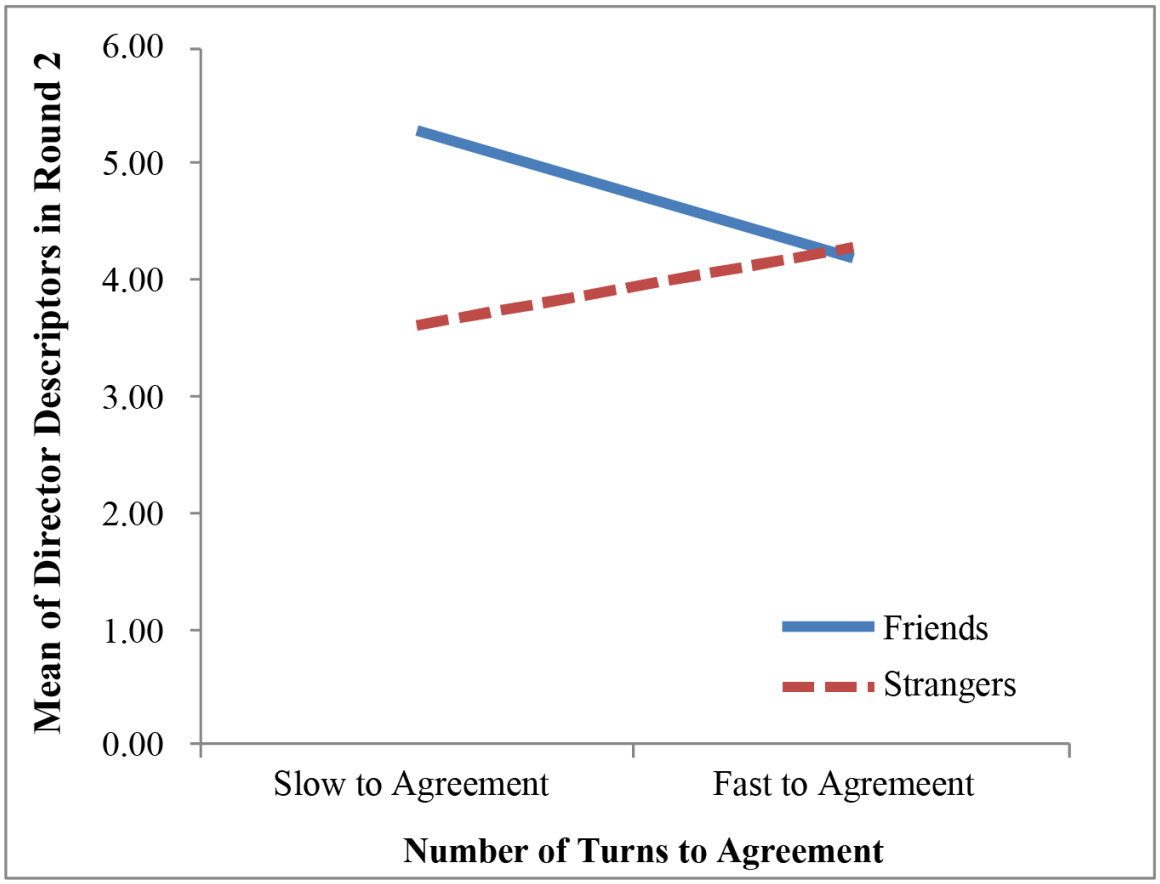

Figure 7. Simple slopes graph for friend and stranger pairs for round 2 director descriptors on round 1 turns.

Overall, there was no evidence of a difference between friends ($M = 4.34$, $SD = 1.91$) and strangers ($M = 3.99$, $SD = 1.64$) in the number of director descriptors used in round 2, $t(46) = -0.67$, $p = .51$, 95% CI [-1.38, 0.69]. A simple slopes analysis of the interaction between round 1 turns and acquaintanceship (Figure 7) was conducted. We defined one standard deviation above the average number of turns as s*low* and one standard deviation below the average number of turns as *fast*. This analysis indicated that the stranger slope did not deviate from zero, stranger slope = -2.19, $t(46) = -1.05$, $p = .30$, but the friend slope did, friend slope = 7.53, $t(46) = 2.61$, $p = .01$, 95% CI [1.74, 13.32]. This indicates that the effect of the number of turns in round 1 on the number of

descriptors used in round 2 was moderated by acquaintanceship. For friend pairs, the fewer turns used to reach agreement on a target in round 1, the fewer director descriptors were needed in round 2. For the strangers, the number of descriptors used in round 2 did not change depending on whether they took fewer or more turns in round 1. See Table 9.

Table 9. Regression table for Analysis 3: Round 2 Director Descriptors. *Sig at .05 level; **Sig. at .01 level; ***Sig at .001 level.

	Model C		**Model D**		**Model E**	
	B	**SE**	**B**	**SE**	**B**	**SE**
Acquaintance	-0.30	0.52	-0.74	0.53	-0.25	0.49
Director (Dir) Extraversion	0.00	0.10	-0.47	0.25		
Acquaintance * Dir Extraversion	0.11	0.14	1.12*	0.43		
R1 Dir Descriptors			0.04	0.09		
R1 Dir Descriptors * Dir Extraversion			-0.09	0.16		
R1 Dir Descriptors * Acquaintance			0.11*	0.05		
Round 1 Turns					7.53**	2.88
Round 1 Turns * Acquaintance					-9.72***	3.55
Adjusted R^2	-0.03		0.097		0.104	
F for Model	0.49		1.84		2.83*	

3.3.1 Analysis 3 Discussion

In analysis 3 we found no evidence of a difference in how many round 2 descriptors strangers used, based on whether they were fast or slow to agree upon a referring expression for a referent in round 1. However, speed at agreeing did make a difference for friend pairs: Directors in friend pairs who had taken more turns to negotiate a referring expression or agree on a referent in round 1 tended to use more descriptors in round 2. Said another way, stranger directors did not adjust their strategy based on how the pair did in round 1, but friend directors did. That being said, the results for analysis 3 should be interpreted with caution as the replacement of round 1 director descriptors

with round 1 turns was done to explore whether any basic measure of how participants performed in round 1 influenced the number of descriptors used in round 2.

4 General Discussion

In line with decades of research using laboratory-based referential communication tasks such as the Tangram Task, we show that the shortening of referring expressions also happens in more naturalistic conversational settings: Directors used fewer descriptors when referring to targets over two rounds of a remote Skype-to-cell-phone referential communication task while their matcher partners had to navigate real-life obstacles on busy streets.

Taking no further variables into account, there was no evidence of a difference between the performance of friends and strangers, nor was there evidence that the number of descriptors used changed among people of varying extraversion levels. Alone, friends and strangers were equally efficient, as were introverts and extraverts. However, acquaintanceship status and extraversion together did affect how quickly pairs could mutually agree upon a referring expression. The more introverted the stranger director, the fewer descriptors were used; that is, for strangers, introversion led to increased efficiency of communication. Friends, however, were not influenced by extraversion. In other words, extraverts and introverts were only distinguishable in stranger pairs.

For both friends and strangers, the influence of extraversion disappeared by the second round. Results of exploratory analyses potentially suggest that existing friendship affected how efficiently pairs performed in the task, as was observed in the interaction between acquaintanceship and the number of turns used in round 1. This time, the effect was observed with friends: Directors in friend pairs were sensitive to whether there were a greater or fewer number of turns before agreeing upon the target in the first round and adjusted their behavior accordingly. The more turns they took in the first round, the more descriptors they used in the second. This trend was not significant for strangers. Whereas friends seemed to adapt to their partners and slow down to accommodate potential difficulties in communication, strangers did not.

As with any task completed in relatively uncontrolled circumstances, there are a number of variables that we did not test in our design that may have affected results. For example, the mobile phone-to-Skype communication in our study may have been constrained by the quality of the audio, the quality of the cellular connection, or even the participant's comfort with using a phone that was not their own. These constraints may have caused participants to repeat themselves and each other more than usual, to speak more loudly to overcome ambient noise from the environment, or to express information differently because they feared getting disconnected. Personality factors may also manifest differently with phone-to-Skype communication versus other communication. For example, extraverts and introverts have different opinions on what constitutes polite mobile phone behavior in public spaces (Love & Kewley, 2005). Consequently, introverted matchers might have designed utterances differently, knowing that bystanders might overhear their conversation.

Differences in task-oriented conversation between friends and strangers are not as evident as people might assume. We provide some evidence that friends and strangers can differ in conversations that are focused on a specific collaborative goal, though their verbal behavior is moderated by extraversion and, potentially, by the conversations they have had in the recent past. Though the in-the-wild method introduces literal and statistical noise, putting people into more naturalistic contexts and examining discourse between interlocutors who have various levels of acquaintanceship can reveal differences that are hidden or discouraged in the laboratory.

Acknowledgements

This research was supported by NSF Grant IIS # 1044693 from the Robust Intelligence Program. We thank our many research assistants who aided in data collection and coding, with special thanks to Haley Biesemeier, Alea Casanova, Ericka Elphick, Andrea Estrada, Steven Ethington, Steven Finet, Jennifer Fong, Beth Hart, Bronwyn Hassall, Erin Hiscock-Wagner, Megan Kostecka, Emilie Kovalik, Tyler Pilgrim, Natalie Serourian, Sean Tang, Jason Tharp, Yana Ulitsky, Elizabeth Williams, and Evelyn Yap. We thank Natalia Blackwell and Lena Reed for contributions to this project. We thank Barbara Di Eugenio and four anonymous reviewers for comments on an earlier version of this manuscript. Correspondence can be addressed to Jean E. Fox Tree (foxtree@ucsc.edu) in the Psychology Department, University of California Santa Cruz, Santa Cruz, CA, 95064.

References

Albright, L., Cohen, A. I., Malloy, T. E., Christ, T., & Bromgard, G. (2004). Judgments of communicative intent in conversation. *Journal of Experimental Social Psychology, 40*, 290-302.

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J. McAllister, J., Miller, J., Sotillo, C., Thompson, H. S. & Weinert, R. (1991). The HCRC map task corpus. *Language and Speech, 34*(4), 351-366.

Andersson, J., & Rönnberg, J. (1995). Recall suffers from collaboration: Joint recall effects of friendship and task complexity. *Applied Cognitive Psychology, 9*(3), 199-211.

Bard, E., Aylett, M., & Lickley, R. (2002). Towards a psycholinguistics of dialogue: Defining reaction time and error rate in a dialogue corpus. In *Proceedings of the Sixth Workshop on the Semantics and Pragmatics of Dialogue* (EDILOG 2002).

Beauchet, O., Dubost, V., Herrmann, F. R., & Kressig, R. W. (2005). Stride-to-stride variability while backward counting among healthy young adults. *Journal of Neuroengineering and Rehabilitation, 2*(1), 26.

Beukeboom, C. J., Tanis, M., & Vermeulen, I. E. (2013). The language of extraversion: Extraverted people talk more abstractly, introverts are more concrete. *Journal of Language and Social Psychology, 32*(2), 191-201.

Blackwell, N., & Fox Tree, J. E. (2012). Social factors affect quotative choice. *Journal of Pragmatics, 44*(10), 1150-1162.

Bortfeld, H., Leon, S. E., Bloom, J. E., Schober, M. F., & Brennan, S. (2001). Disfluency rates in spontaneous speech: Effects of age, relationship, topic, role, and gender. *Language and Speech, 44*(2), 123-149.

Branigan, H., Lickley, R., & McKelvie, D. (1999). Non-linguistic influences on rates of disfluency in spontaneous speech. In *Proceedings of the 14th International Conference of Phonetic Sciences*.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory & Cognition, 22*(6), 482-493.

Brennan, S. E., Schuhmann, K. S., & Batres, K. M. (2013). Entrainment on the move and in the lab: The walking around corpus. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*.

Campbell, A., & Rushton, J. P. (1978). Bodily communication and personality. *British Journal of Social and Clinical Psychology, 17*(1), 31-36.

Carment, D., Miles, C., & Cervin, V. (1965). Persuasiveness and persuasibility as related to intelligence and extraversion. *British Journal of Social and Clinical Psychology, 4*(1), 1-7.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22*, 1-39.

Clark, H. H. (1996). *Using language*. Cambridge University Press.
Cohen, J., Cohen, P., West, S., & Aiken, L. (2003). *Applied multiple regression/correlation analysis for the social sciences* (3rd edition). Lawrence Erlbaum Associates.
Cuperman, R., & Ickes, W. (2009). Big Five predictors of behavior and perceptions in initial dyadic interactions: Personality similarity helps extraverts and introverts, but hurts "disagreeables". *Journal of Personality and Social Psychology, 97*(4), 667-684.
Dewaele, J. M., & Furnham, A. (1999). Extraversion: The unloved variable in applied linguistic research. *Language Learning, 49*(3), 509-544.
Dunne, M., & Ng, S. H. (1994). Simultaneous speech in small group conversation: All-together-now and one-at-a-time? *Journal of Language and Social Psychology, 13*(1), 45-71.
Feldstein, S., & Sloan, B. (1984). Actual and stereotyped speech tempos of extraverts and introverts. *Journal of Personality, 52*(2), 188-204.
Fleming, J. H., Darley, J. M., Hilton, J. L., & Kojetin, B. A. (1990). Multiple audience problem: A strategic communication perspective on social perception. *Journal of Personality and Social Psychology, 58*(4), 593-609.
Fox Tree, J. E. (1999). Listening in on monologues and dialogues. *Discourse Processes*, *27*(1), 35-53.
Fox Tree, J. E. (2007). Folk notions of *um* and *uh*, *you know*, and *like*. *Text & Talk, 27*(3), 297-314.
Fox Tree, J. E., & Clark, N. B. (2013). Communicative effectiveness of written versus spoken feedback. *Discourse Processes*, *50*(5), 339-359.
Fox Tree, J. E., Mayer, S. A., & Betts, T. E. (2011). Grounding in instant messaging. *Journal of Educational Computing Research, 45*(4), 455-475.
Froming, W. J., & Carver, C. S. (1981). Divergent influences of private and public self-consciousness in a compliance paradigm. *Journal of Research in Personality, 15*(2), 159-171.
Fussell, S. R., & Krauss, R. M. (1989). Understanding friends and strangers: The effects of audience design on message comprehension. *European Journal of Social Psychology*, *19*(6), 509-525.
Gifford, R., & Hine, D. W. (1994). The role of verbal behavior in the encoding and decoding of interpersonal dispositions. *Journal of Research in Personality, 28*(2), 115-132.
Gill, A. J., & Oberlander, J. (2003). Perception of e-mail personality at zero-acquaintance: Extraversion takes care of itself; neuroticism is a worry. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*.
Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality, 37*(6), 504-528.
Gould, O. N., Osborn, C., Krein, H., & Mortenson, M. (2002). Collaborative recall in married and unacquainted dyads. *International Journal of Behavioral Development, 26*(1), 36-44.
Guydish, A., D'Arcey, J. T., & Fox Tree, J. E. (2020). Reciprocity in conversation. *Language and Speech.* Advance online publication. doi: 10.1177/0023830920972742
Haxby, J. V., Parasuraman, R., Lalonde, F., & Abboud, H. (1993). SuperLab: General-purpose Macintosh software for human experimental psychology and psychological testing. *Behavior Research Methods, Instruments, & Computers*, *25*(3), 400-405.
Heylighen, F., & Dewaele, J.-M. (1999). Formality of language: Definition, measurement and behavioral determinants. *Interner Bericht, Center "Leo Apostel", Vrije Universiteit Brüssel*, 4.
Hill, A., Bohil, C., Lewis, J., & Neider, M. (2013). Prefrontal cortex activity during walking while multitasking: An fNIR study. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*.
Hope, L., Ost, J., Gabbert, F., Healey, S., & Lenton, E. (2008). "With a little help from my friends…": The role of co-witness relationship in susceptibility to misinformation. *Acta Psychologica, 127*(2), 476-484.
Hornstein, G. A. (1985). Intimacy in conversational style as a function of the degree of closeness between members of a dyad. *Journal of Personality and Social Psychology, 49*(3), 671-681.

Horton, W. S., & Spieler, D. H. (2007). Age-related differences in communication and audience design. *Psychology and Aging, 22*(2), 281-290.

Isaacs, E. A., & Clark, H. H. (1987). References in conversations between experts and novices. *Journal of Experimental Psychology: General, 116*(1), 26-37.

Jucker, A. H., & Smith, S. W. (1998). And people just you know like "wow": Discourse markers as negotiating strategies. In Andreas H. Jucker & Yael Ziv (Eds.), *Discourse markers: Description and theory*, pp. 171–201. John Benjamins.

Jung, J., Lee, Y., & Karsten, R. (2012). The moderating effect of extraversion–introversion differences on group idea generation performance. *Small Group Research*, *43*(1), 30–49.

Kemper, S., Othick, M., Warren, J., Gubarchuk, J., & Gerhing, H. (1996). Facilitating older adults' performance on a referential communication task through speech accommodations. *Aging, Neuropsychology, and Cognition, 3*(1), 37-55.

Kenny, D. A., & Acitelli, L. K. (2001). Accuracy and bias in the perception of the partner in a close relationship. *Journal of Personality and Social Psychology; Journal of Personality and Social Psychology, 80*(3), 439-448.

Levesque, M. J., & Kenny, D. A. (1993). Accuracy of behavioral predictions at zero acquaintance: A social relations analysis. *Journal of Personality and Social Psychology, 65*(6), 1178-1187.

Li, B., Zhou, H., He, J., Wang, M., Yang, Y., & Li, L. (2020, November). On the sentence embeddings from BERT for semantic textual similarity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9119-9130.

Liu, K., Fox Tree, J. E., & Walker, L. (2016). Coordinating communication in the wild: The Artwalk dialogue corpus of pedestrian navigation and mobile referential communication. In *Proceedings of the International Conference on Language Resources and Evaluation*, Portorož, Slovenia, pp. 3159-3166.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692

Love, S., & Kewley, J. (2005). Does personality affect peoples' attitude towards mobile phone use in public places? *Mobile Communications* (pp. 273-284). Springer.

Love, S., & Perry, M. (2004). Dealing with mobile conversations in public places: Some implications for the design of socially intrusive technologies. In *CHI'04 Extended Abstracts on Human Factors in Computing Systems*.

Neider, M. B., McCarley, J. S., Crowell, J. A., Kaczmarski, H., & Kramer, A. F. (2010). Pedestrians, vehicles, and cell phones. *Accident Analysis & Prevention, 42*(2), 589-594.

Nelson, P. A., Thorne, A., & Shapiro, L. A. (2011). I'm outgoing and she's reserved: The reciprocal dynamics of personality in close friendships in young adulthood. *Journal of Personality, 79*(5), 1113-1148.

Oberlander, J., & Gill, A. J. (2006). Language with character: A stratified corpus comparison of individual differences in e-mail communication. *Discourse Processes, 42*(3), 239-270.

Oulasvirta, A., Tamminen, S., Roto, V., & Kuorelahti, J. (2005). Interaction in 4-second bursts: The fragmented nature of attentional resources in mobile HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Peeters, M. A., Van Tuijl, H. F., Rutte, C. G., & Reymen, I. M. (2006). Personality and team performance: a meta-analysis. *European Journal of Personality, 20*(5), 377-396.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology, 77*(6), 1296-1312.

Planalp, S., & Benson, A. (1992). Friends' and acquaintances' conversations I: Perceived differences. *Journal of Social and Personal Relationships, 9*, 483-506.

Preacher, K. J., Curran, P. J., & Bauer, D. J. (2003). Simple Intercepts, Simple Slopes, and Regions of Significance in LCA 2-Way Interactions. Retrieved from http://www.quantpsy.org/interact/lca2.htm

Rawlings, D., & Carnie, D. (1989). The interaction of EPQ extraversion with WAIS subtest performance under timed and untimed conditions. *Personality and Individual Differences, 10*(4), 453-458.

Reimers, N. & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3973-3983.

Savitsky, K., Keysar, B., Epley, N., Carter, T., & Swanson, A. (2011). The closeness-communication bias: Increased egocentrism among friends versus strangers. *Journal of Experimental Social Psychology, 47*(1), 269-273.

Scherer, K. R. (1978). Personality inference from voice quality: The loud voice of extroversion. *European Journal of Social Psychology, 8*(4), 467-487.

Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology, 21*(2), 211-232.

Smoski, M., & Bachoroski, J.-A. (2003). Antiphonal laughter between friends and strangers. *Cognition & Emotion, 17*(2), 327-340.

Speed, L. J., Wnuk, E., & Majid, A. (2016). Studying psycholinguistics out of the lab. *Research methods in psycholinguistics and the neurobiology of language*. Wiley-Blackwell.

Stinson, L., & Ickes, W. (1992). Empathic accuracy in the interactions of male friends versus male strangers. *Journal of Personality and Social Psychology, 62*(5), 787-797.

Thorne, A. (1987). The press of personality: A study of conversations between introverts and extraverts. *Journal of Personality and Social Psychology, 53*(4), 718- 726.

Tolins, J., Zeamer, C., & Fox Tree, J. E. (2018). Overhearing dialogues and monologues: How does entrainment lead to more comprehensible referring expressions? *Discourse Processes*, *55*(7), 545-565.

Tolins, J., Liu, K., Neff, M., Walker, M., & Fox Tree, J. E. (2016). A verbal and gestural corpus of story retellings to an expressive embodied virtual character. In *Proceedings of the International Conference on Language Resources and Evaluation*, Portorož, Slovenia, pp. 3461-3468.

Truong, K. P., & Heylen, D. (2012). Measuring prosodic alignment in cooperative task-based conversations. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Truong, K. P., & Trouvain, J. (2012). Laughter annotations in conversational speech corpora-possibilities and limitations for phonetic analysis. In *Proceedings of the 4th International Workshop on Corpora for Research on Emotion Sentiment and Social Signals*, 20-24.

Uziel, L. (2007). Individual differences in the social facilitation effect: A review and meta-analysis. *Journal of Research in Personality, 41*(3), 579-601.

Wang, H.-C., Fussell, S. F., & Setlock, L. D. (2009). Cultural difference and adaptation of communication styles in computer-mediated group brainstorming. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Wu, S., & Keysar, B. (2007). The effect of culture on perspective taking. *Psychological Science, 18*(7), 600-606.

Xu, Y. (2010). In defense of lab speech. *Journal of Phonetics, 38*(3), 329-336.

Yogev-Seligmann, G., Hausdorff, J. M., & Giladi, N. (2008). The role of executive function and attention in gait. *Movement Disorders, 23*(3), 329-342.

Appendix A
Descriptor Matching Instructions

The goal of this coding is to see whether participants used the same terms to describe objects to each other (*lexical entrainment*). For that reason, categorize descriptors as non-matches if unclear. Use the following guides to match first. Dashed words count as multiple words. Note that most of the descriptors in this set will be single words. Descriptors are always separated by commas. See the end of this document for examples.

Strict Way

Single words. For each of the director descriptors, see if it exists in the matcher descriptors. If it is an exact match, count it as a match. In case of a close or partial match, determine whether the meaning is similar enough to be treated as a match: (1) Is the root of the word the same (e.g., yellow and yellowish are the same root word, flying and flies are the same root word)? If so, consider it a match. (2) Otherwise, it's not a match.

Multiple words. (1) Are all of the content words identical? (e.g., 'surrounded by plants' and 'plants are surrounding it' have two identical content words, so it is a match. (2) If the content words are identical but one is *negated* (e.g., 'holding hands' and 'not holding hands') it is *not* a match. (3) Otherwise, it's not a match.

Mix of single and multiple words. Do the content words all match? (e.g., "The baseball" and "baseball" match, but "white baseball" and "baseball" do not match).

Flexible Way

Single words. (1) Is the word almost identical in meaning (e.g., tan, khaki, beige)? If so, consider it a match. (2) Otherwise, it's not a match.

Multiple words. (1) Are most of the content words at least similar? (e.g., 'it's kind of round and a dark blue purple color' and 'it's circular and purpleish') If you can get the same general idea, consider it a match. (2) If one description is more specific than the other/the verbiage is different but related to the same object (e.g., 'it has a baseball' versus 'it's holding a baseball'), it can be a match. (3) Otherwise, it's not a match.

Example

Director (Number of descriptors = 17)

painting, on wall, or on sidewalk, jaggedy looking, turquoise, two, little, purple, egg-shaped figures, little, black legs, black arms, sticking out, no head, tiny arms and legs, at the bottom, brown underneath the turquoise

Matcher (Number of descriptors = 11)

mural, like a rainbow, boat, turquoise, green strip, reddish brown strip, bar, with bunch of hieroglyphs, little figures, strip of blue, and green

Matches the strict way

turquoise

Matches the flexible way

turquoise/turquoise; mural/painting; egg-shaped figures/little figures