See What I'm Saying? Using Dyadic Mobile Eye Tracking to Study Collaborative Reference

Darren Gergle[*†] **and Alan T. Clark**[*]
[*]Dept. of Communication Studies, [†]Dept. of Electrical Engineering and Computer Science
Northwestern University
Evanston, IL 60208 USA
{dgergle, alan-clark}@northwestern.edu

ABSTRACT
To create intelligent collaborative systems able to anticipate and react appropriately to users' needs and actions, it is crucial to develop a detailed understanding of the process of collaborative reference. We developed a dyadic eye tracking methodology and metrics for studying the multimodal process of reference, and applied these techniques in an experiment using a naturalistic conversation elicitation task. We found systematic differences in linguistic and visual coordination between pairs of mobile and seated participants. Our results detail measurable interactions between referential form, gaze, and spatial context and can be used to enable the development of more natural collaborative user interfaces.

Author Keyword
Dual eye tracking, gaze, reference, language, dyadic interaction, shared visual space

ACM Classification Keywords
H.5.3 [Information Interfaces and Presentation]: Group and Organization Interfaces – collaborative computing, computer-supported cooperative work, computer-mediated communication

General Terms
Human Factors, Experimentation, Measurement, Design

INTRODUCTION
A great deal has been made about recent technological trends that promise to move us from the era of mouse and keyboard to a new world of natural and conversational user interfaces. Popular accounts pledge systems that, like a good personal assistant, anticipate users' needs and actions. By using machine perception and natural language processing, these interfaces will intelligently assess language use, understand what in the environment people are talking about, and react accordingly. Such interfaces, it is suggested, will result in an effortless and natural interaction paradigm for the user.

Yet, it is a monumental task to understand how humans interact naturally in contextualized physical environments, let alone build machines that can do the same. If natural user interfaces are going to deliver on their promise, we first need to better understand natural human interactions.

One particular area of natural interaction that is essential to communication, coordination, and collaboration is that of reference. Reference is how we specify the particular person, object or entity that we are talking about [9]. By examining reference in a dynamic mobile scenario where collaborative pairs can freely move about an environment, we answer a call to better understand the coordinating role of reference in CSCW environments, addressing an area of research previously described as "unexplicated and under-explored" [21]. Our findings detail measureable patterns that exist between referential form, gaze, and spatial context, and can be used to enable the development of more natural and collaborative user interfaces.

The goal of this work is to advance the knowledge necessary to develop and deploy successful conversational systems that interact with humans during collaborative physical activities, serve as the basis for human-agent and human-robot interactions, and support conversational systems that dynamically adapt based on predictive models of group behavior in natural environments. We aim to do this by developing a more detailed, formal and complete theoretical account of the coordination dynamics that take place in collaborative physical environments, and further uncovering the link between language and physical actions that serve collaboration [18].

This research contributes: (1) an advanced understanding of the dynamics of reference as they take place in collaborative pairs freely moving about in an unconstrained physical space, (2) a mobile gaze tracking system for studying collaboration and dyadic gaze patterns that is able to automatically measure gaze to real-world objects, and (3) a new set of metrics and processes for analyzing and using these eye tracking data.

CSCW 2011, March 19–23, 2011, Hangzhou, China.

LITERATURE REVIEW

The basic communicative act of talking about people, objects and events requires a context-dependent understanding of reference. Referential expressions, such as "I," "here," "that," or even "tomorrow" have various target referents depending on the speaker, the hearer, the physical domain, the time, or the current context. Two speakers may use the same referential expression in an identical sentence, but it is the nature of indexical reference that they may be referring to completely different things (e.g., my "here" may be different than your "here" if we are standing at opposite ends of a room).

The English language provides a number of ways to refer to things. For example, an object referred to using the expression, "it looks kind of like a merry-go-round," may subsequently be referenced using a variety of forms such as: the merry-go-round, this, that, it, the small green and yellow thing, etc. The use of different expressions varies with the accessibility[1] of the object in the pair's current discourse [2, 20]. Unfamiliar objects typically receive the most detailed description (e.g., "the little brown building"), while referents that have already been introduced require less detailed referential forms (e.g., "the building," "that one," "it"). In short, successful reference takes into account the common ground between speakers [14].

Violations of referential conventions can result in ambiguity, confusion and incomprehensible utterances. For example, it would be infelicitous for a speaker to refer to "the merry-go-round" using the pronoun "it" if she had since advanced the discourse by discussing several other architectural structures. Likewise, she should only include additional modifiers such as "the small green and yellow merry-go-round" if she believes the shorter reference is ambiguous, such as when there are two visible objects resembling merry-go-rounds.

Multimodal Reference

Reference research has, until quite recently, primarily resided within the domain of linguistics and pragmatics. Yet, as noted by Hindmarsh & Heath [21], certain references can only be understood when the surrounding physical context is taken into account. Researchers have begun to explore the fact that objects for conversation are evoked though multiple avenues: language, action, movement, or other elements of the pragmatic context. This break from a language-only representation of reference brings about notions of situationally-evoked referents [35], visually-salient entities [12, 19, 22], or, more generally, a system of embodied reference [21]. A common thread among these approaches is that visual cues are combined with linguistic cues to enable effective reference, and that our technological systems must take them both into account [19].

A related area of work concerns dyadic gaze and language use [1]. Most computational studies of gaze and language use have focused on higher-level discourse properties such as turn-taking and question-asking [33], comprehension [30], storytelling [24], communicative engagement [31], or task properties [17]. However, a few studies have looked at the relationship between eye gaze and referential form. Nakano and colleagues [25] explored the relation between various dialogue acts and non-verbal behaviors and showed that speakers look at their partners in order to ground references to new entities. Bard and colleagues [4] focused on a 2D desktop display and found that mutual gaze to objects is not always high during reference yet it is significantly above chance. They also showed that the accessibility of referential forms is temporally tied to gaze coordination.

While previous work begins to elucidate a rich variety of pragmatic and contextual factors that can influence referential form, the studies typically examine reference within relatively static referential domains (for a notable exception see [21]). Yet, in natural everyday conversation interlocutors move about, shift their body positions, and gaze in different directions. As a result, attentional cues and the pair's referential domain (i.e., the set of objects a given reference might refer to) dynamically shifts throughout the course of a conversation. Such dynamics do not appear in typical experimental settings that artificially constrain the space, resulting in an incomplete picture of the ways in which shifting spatial context influences reference.

Reference and Technology

Contrary to the emerging view of reference as a multimodal and embodied construct, engineering practice has long espoused a simple conduit model of communication that consists of packaging or encoding information in a message, transmitting the message through some channel or medium, and decoding of the content by a receiver. This effectively reduces communication to simple message passing, quantizing collaboration into convenient packages for designers and engineers [notable exceptions include 16, 25, 34].

A richer understanding of referential behavior is crucial to the development of the next wave of intelligent collaborative user interfaces. Emerging technologies ranging from conversational agents that interact directly with humans on collaborative physical tasks [11], to VR and AR systems that attempt to comprehend spoken references to objects in the environment [8, 28], to video-mediated communication systems that track conversation and automatically adapt their views based on what the pairs need to see [26, 29], would benefit from more advanced computational models of human referring behavior and an understanding of the ways in which context influences

[1] Accessibility is based on the idea that references "instruct addressees to retrieve a certain piece of [g]iven information from memory" [2, p.29]. More accessible referents can be understood as being on people's minds, whereas less accessible referents would be more difficult to retrieve from memory.

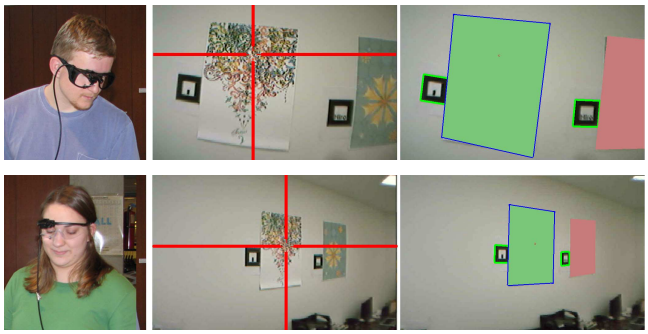


Figure 1. The dual mobile eye tracking system converts raw eye tracking output (middle images) into a form that presents what a person is looking at (green objects in right images) and what other objects are in the field of view (pink objects).

collaborative reference. Furthermore, new technologies that provide lightweight mobile eye tracking capabilities [7] are quickly becoming available as platforms for collaborative technologies that reside in everyday physical settings away from the desktop.

As we begin to develop systems and applications based on these platforms and incorporate computational models of collaborative reference, we will need to answer several questions such as: How are the generation and comprehension of referring expressions influenced by the availability of various features of a shared visual context? How are these features crucial to the coordination processes that serve successful collaboration, and how can one model them in a way that can serve the development of more successful collaboration technologies?

To begin addressing these challenges, we undertook a study of reference as it takes place with mobile pairs able to freely move about a physical environment and compare their referential dynamics to pairs in a static seated environment. This approach allows us to understand and make use of the interrelated coordination mechanisms of spoken discourse context (e.g., whether an entity has been mentioned), gaze patterns (e.g., whether the addressee is looking at the speaker's intended referent), and the spatial context in which collaboration takes place (e.g., whether moving and stationary pairs use deictic references differently).

OUR APPROACH

A serious constraint to furthering our understanding of multimodal reference stems from a lack of tools and methodologies to gather the rich process data needed, and a reliance upon rigid experimental paradigms that often consist of one-shot communication events in constrained task environments [for a critique see 10].

To address these concerns, we developed a new experimental methodology: a real-time, naturalistic, dyadic eye tracking approach with a set of metrics that can be used to study collaboration. This approach is unique in that it collects a fine-grained process record from two people, in real time, as they interact in a naturalistic environment. Importantly, our system is able to automatically recognize gaze to objects in the physical space.

Related Methodologies

Single user eye tracking methods have been used to study multimodal communication for several decades [15, 32], but recent developments in wearable eye tracking systems now permit the collection of eye movement data in more natural and unconstrained environments. Studies using this approach have explored individual physical tasks such as meal preparation [23] and hand washing [27]. A recent development concerns the use of *dyadic* eye tracking approaches for understanding gaze coordination [4, 13, 30]. Dyadic eye tracking shifts the interpretation from a cognitive focus (e.g., What is John thinking when he looks at the screen?) to a more social one (e.g., Does Susan's gaze pattern influence John's speech?) (P. Dillenbourg, personal communication, June 2009). However, these studies typically examine highly constrained interactions. While the mobile single user approaches permit the study of natural behavior, the latter dyadic approaches demonstrate how coordination and collaboration can be examined.

A Novel Dyadic Eye Tracking Methodology

We combine the two approaches into a new dyadic eye tracking methodology. Our approach makes use of two synchronized mobile eye trackers that gather eye movements and gaze patterns in an unconstrained physical environment. Figure 1 illustrates the system capturing real-time data from two people at the same time. The middle frame illustrates the post-processed data from the Applied Science Laboratories (ASL) EyeVision system (http://asleyetracking.com/site/), and the frames on the right demonstrate how this information can be coupled with a computer vision module to extract a running log of gaze patterns to tagged elements in the physical environment. While we currently use a marker-based vision approach, any computer vision technique can be applied to demarcate regions of interest in physical space.

Figure 2 diagrams the architecture for interfacing a pair of ASL Mobile Eye units with our custom computer vision modules developed using the ARToolkit (http://www.hitl.washington.edu/artoolkit/) and OpenCV libraries (http://opencv.willowgarage.com/wiki). This system collects an automated, fine-grained temporal record of real-time gaze patterns and speech from two different people in an untethered physical environment, allowing us to analyze the process of multimodal coordination in real-world physical contexts.

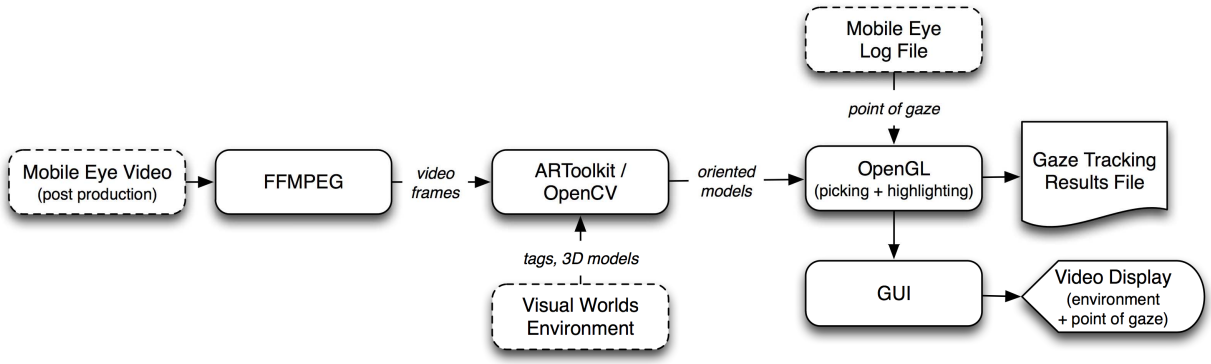


Figure 2. Architecture for integrating dual eye tracking and automatic object recognition (previously existing components are in dashed elements while our new system components are in solid elements).

Behavioral Evaluation Metrics

Several data processing and analytical challenges need to be met for this approach to be useful. Repeated measures drawn from pairs whose performance is contingent upon one another increases the likelihood that the data are correlated and non-independent. We addressed this by analyzing gaze overlap using generalized linear mixed-models regression with covariance modeling to account for the lack of independence between measurements [5].

Another challenge is the non-linear form of the data that results from shifting patterns of gaze overlap, with speakers' and addressees' gaze coordination often occurring at an offset of several hundred milliseconds. To account for this shifting lag in gaze coordination, the 'leader' and 'follower' need to be identified in order to make gaze overlap a strong indicator of coordination of attention. Without such corrections, established measures such as cross-recurrence [6] can become noisy. Our analysis uses detectable discourse markers (i.e., current speaker, recent referential history) to model initiative and incorporate user-defined directional temporal lags into gaze metrics (for a related approach see [4]).

Our evaluation of gaze overlap is linked with the dyad's referential state, a critical distinction of our approach. We are able to differentiate between gaze overlap that is occurring generally (and by chance) and gaze overlap that reflects targeted attention to the current object by determining the intended referent for referring expressions in our corpus. This permits a more sophisticated analysis of the dynamic role of gaze in attention and conversation than a non-contextual gaze overlap or cross-recurrence measure. Our approach also allows us to analyze when specific attention dynamics are at play in conversation, such as instances where there is above-chance gaze overlap to objects that are not currently the discourse focus, or instances where the addressee is looking at the speaker's intended referent at below-chance levels.

HYPOTHESES

As previously discussed, the research literature suggests numerous factors involved in referring behavior as it takes place in natural environments. Based on this background we posit a number of hypotheses regarding referential form, spatial context, gaze coordination, and their interactions.

Spatial Context and Referential Form

Consistent with accessibility theory [2], pairs should produce elongated referential forms (e.g., a deictic pronoun such as "this" or "those" with numerous modifiers) when first mentioning an object or shifting the discussion to a new referent. For continuing discussion of referents, a more accessible linguistic form such as a pronoun is expected. Hence,

H1: *Pairs will use more elongated (i.e., less accessible) referential forms for referential initiations and shifts between referents, and less elongated (i.e., more accessible) forms for continued references to the same object.*

However, given the relationship between reference and physical space, we expect referential forms to reflect differences in spatial context. Users able to move through a shared space will be able to use their physical positioning as a form of dynamic "visual conduct" [21]. We expect that mobile users will be less reliant on linguistic detail to initially pick out referents and direct attention, using their movement towards objects as a coordination mechanism that makes referents more accessible. This should be reflected in pairs' use of deictic (pointing) terms such as demonstratives – "this", "that", "these", and "those". In particular, there should be different patterns of spatially marked local or remote demonstratives (e.g., "this" marks a referent as local). Thus, we expect:

H2a: *Mobile pairs will use local deictic demonstratives to refer to objects more often than will seated pairs.*

H2b: *This distinction will matter more during referential initiations and shifts between referents and will dissipate over time due to the increased role of spoken discourse context.*

Similarly, if we expect the mobile pairs to use their movement to supplant linguistic coordination mechanisms, then the inverse should occur with respect to remote deictic demonstratives (i.e., "that"):

H3a: *Seated pairs will use remote deictic demonstratives to refer to objects more often than will mobile pairs.*

Figure 3. Experimental conditions from left to right: Seated (side-by-side), Seated (across), and Mobile.

H3b: *This distinction will matter more during referential initiations and shifts between referents and will dissipate over time due to the increased role of spoken discourse context.*

In the instances when speakers add additional disambiguating information to their deictic references, the type of information they add will reflect the availability (or lack thereof) of movement as a spatial coordination mechanism. Bangerter [3] noted that the production of feature (e.g., "tall and red") and locative (e.g., "on the left") information increased with distance from a spatial referent. Because mobile pairs can reposition themselves, we predict they will require fewer locative terms to direct attention to referents than the seated pairs.

H4: *Seated pairs will include locative information in their deictic references more often than will mobile pairs.*

Alternatively, mobile pairs will use more descriptive terms if they need additional information to guide attention to a referent.

H5: *Mobile pairs will include feature information in their deictic references more often than will seated pairs.*

Gaze Coordination

Previous studies have demonstrated the role of dyadic gaze coordination in conversation. However, they focused on gaze overlap occurring in asynchronous, remote, or tightly scripted interactions. We predict that gaze coordination is related to reference in a collocated, natural dyadic interaction.

H6: *Pairs' gaze coordination will occur at above-chance levels when referring to objects in their shared space.*

Because seated pairs have fewer alternative coordination mechanisms (e.g., movement or positioning) to rely upon, we expect:

H7: *Seated pairs will exhibit greater gaze overlap than will mobile pairs.*

Gaze Coordination and Referential Form

In addition to using gaze to inform word choice, speakers will monitor their addressees' gaze and select referential forms to direct attention accordingly. We expect that speakers will be more likely to use demonstrative reference (e.g., "this one") when the addressee is not attending to the speaker's planned referent.

H8: *The speaker will be more likely to use a deictic demonstrative if the addressee is not already attending to the referent.*

Spatial Context, Gaze Coordination and Referential Form

Shared gaze is a mechanism that provides for joint focus of attention. However, the ability to evoke referents with physical movement should reduce the need for gaze coordination when making deictic references. Also, as suggested by Nakano et al. [25], we predict decreased gaze overlap when pairs are introducing new referents.

H9: *Mobile pairs will have lower gaze overlap for local demonstratives than will seated pairs.*

However, mobile pairs will still need to rely on gaze coordination for remote references.

H10: *Mobile pairs will have similar levels of gaze overlap as seated pairs for remote demonstratives.*

Similar to H9, when mobile pairs refer to an object they can use their spatial positioning to evoke the referent and look at the other person to monitor understanding. Thus,

H11: *During initial references to an object, mobile pairs' gaze overlap will be lower than seated pairs' gaze overlap.*

THE STUDY

This study aims to explore the interplay of gaze coordination, spatial context, and linguistic detail and form in the process of collaborative reference. We employed a naturalistic conversational elicitation task we have developed—a dyadic "guessing game"—that involves generation and comprehension of descriptions of objects as part of a negotiation about a set of four abstract sculptures. Data were compared across spatial conditions to better understand the role of movement and position in referential behavior. Both raw data and linguistic and spatial data were drawn from head-mounted mobile eye trackers as well as stationary cameras positioned within the experiment space.

METHOD

Participants

66 participants (assigned to 33 pairs) were recruited from the campus of a mid-sized U.S. university. Four pairs were removed from the analysis due to audio recording issues that prohibited accurate transcription.

Procedure

Participants performed a collaborative referential elicitation task. Four Lego™ constructions were built to resemble abstract modern sculpture and participants were told that one of the sculptures was a replica of a real artwork that had recently sold at auction for $100,000. The sculptures were designed to provide referential competition that would require disambiguation. The objects were horizontally symmetrical, with all four sides identical, so participants would not acquire added information about the objects by viewing from different angles. Participants were asked to discuss the objects before jointly producing a final ranking from most to least likely to be the replica. We designed this open-ended conversational task in order to elicit a large amount of unconstrained natural dialogue.

Pairs were randomly assigned to one of three spatial context conditions shown in figure 3. In the seated side-by-side condition, participants sat next to each other approximately 20" (~.51m) apart shoulder-to-shoulder with sculptures on the table in front of them. In the seated across condition, participants sat facing across the table with the objects in between them. In the mobile condition, participants stood and were free to move around as they discussed the objects. The table was ~80" (2.03m) per side with ~36"-96" (~.92-2.44m) of clearance from the wall.

Data Collection and Pre-Processing

We annotated conversation transcripts for each discrete reference to the sculpture objects. Each referring expression was coded for referential form (e.g., indefinite, definite, deictic, pronoun) and descriptive markers (e.g., location or feature information). Two annotators identified the intended referent, using gaze and the conversational context as a guide. Their coding was reliable (Cohen's kappa = .80).

Gaze patterns were recorded using two mobile eye trackers that allowed for free movement of the head and body. The previously described system automatically coded the data for object fixations by each individual and for gaze overlap, which captured the proportion of fixations on an object that overlapped between the speaker and addressee. In order to compensate for the noise and occasional data loss encountered with our mobile eye tracking system, we followed common practice and only analyzed the gaze sampling points in which both participants' trackers were successfully able to calculate point-of-gaze. Although this method excludes blinks and some saccades, it is also more resilient to individual or pair differences in recording quality or physical compatibility with the equipment, compensating for data loss that can skew gaze coordination rates unrealistically low.

Statistical Approach

We used multi-level mixed-effects logistic models to analyze the language results for all binary outcome variables such as deixis (yes/no). The independent variables included spatial condition, reference shift, spatial condition × reference shift, speaker gender, pair gender and degree offset. Because observations were not independent, pair (nested within spatial condition) was modeled as a random effect.

Generalized linear mixed-model regression techniques with covariance modeling were used to analyze the gaze results. Gaze overlap (measured in a ±1500ms window around the onset of the referring expression) was the dependent variable. Spatial condition, remote deixis, local deixis, spatial condition × remote deixis, spatial condition × local deixis, reference order, and reference duration were included as independent variables. Pair (nested within spatial condition) was modeled as a random effect.

The temporal relationship between partner attention and referential form examined the addressee's gaze 2000ms before reference onset. A multi-level mixed-effects logistic model was used with deictic form as a binary dependent variable. Addressee gaze, spatial condition, reference shift, spatial condition × reference shift, reference duration and reference order were independent variables. Pair (nested within spatial condition) was modeled as a random effect.

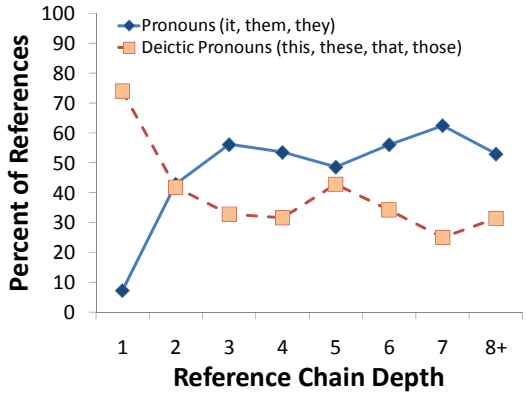


Figure 4. Percent of References by Reference Chain Depth and Reference Type (pronoun or deictic).

RESULTS

We examined 1,473 references drawn from 29 pairs (9 side-by-side, 10 across, 10 mobile). The pairs averaged 50.03 discrete references (SD = 21.82), with little difference in production rates across conditions (M_{free} = 50.5 (SE = 7.18), M_{across} = 49.1 (7.18), M_{side} = 53.0 (7.57)).

Spatial Context and Referential Form

In keeping with Ariel's theory of accessibility [2], reference initiations and reference shifts to different objects exhibited elongated referential forms (e.g., a deictic pronoun with additional modifiers), while references to the same piece were associated more strongly with pronominal forms such as "it" (see figure 4). These findings support H1.

The mobile pairs produced a higher proportion of local deictic references than the seated pairs (see figure 5, top panel). The odds of producing local markings decreased by

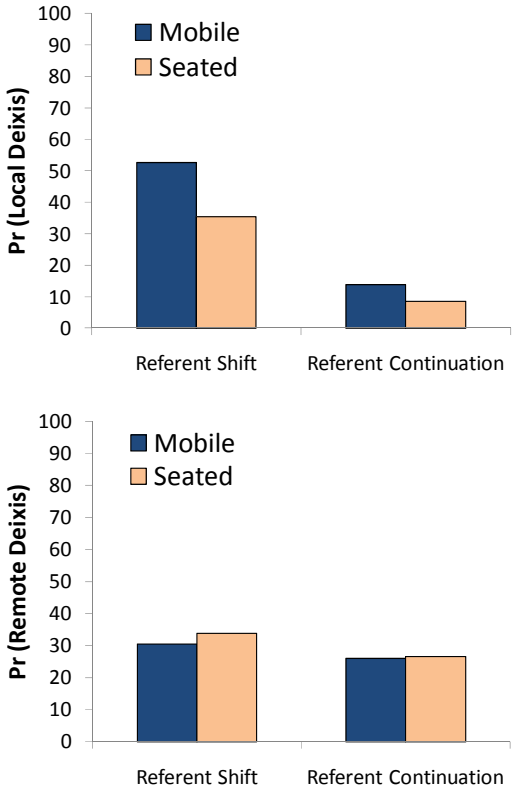


Figure 5. Local Deixis (top) and Remote Deixis (bottom) by mobility condition and referent shift.

50.2% for those in the seated conditions (z = -2.11, p < .05)[2]. In other words, the pairs in the mobile condition appeared to use their body position to evoke potential referents, particularly those sculptural pieces close by, supporting H2a. Also, when all pairs shifted from talking about one sculpture to another they were much more likely to introduce the new referent using a local marking. The odds of generating a local deictic form increased by 5.7 times between reference shifts and initiations (b_{logit} = 1.91; z = 12.24, p < .001). These results support H2b.

However, in contrast to the findings for local deictic references, differences were not found in the use of remote markings (see figure 5, bottom panel). For pairs in the seated conditions, the odds of producing remote markings were not detectably different than those in the mobile conditions (z = 0.46, p = .65). Similarly, while the odds moved in the expected direction when a referential shift occurred, the results were not significant (z = 1.49, p = .136). Thus, we found no support for H3a and H3b.

While these results examine changes in the form of the demonstrative pronoun, another element of reference that can disambiguate is the content of an appended phrase.

When the pairs continued talking about the same sculpture they were much more likely to include modifying information along with the deictic referent (e.g., "this *four-armed thing*"). Stated another way, when a referential shift occurred, the odds of generating a deictic form with additional information decreased by 67% (b_{logit} = -1.11; z = -2.19, p = .028). This result may reflect a preference by the speakers to initially use more efficient referential forms but then revert to more detailed forms if further coordination of attention to the referent is required or grounding on a referential term is still needed or requires further clarification.

While there were no detectable differences between the mobile and seated pairs in the rate of inclusion of additional information (z = -0.60, p = .55), the type of information included was substantially different.

We examined whether the information was locative (i.e., location-based such as "the one to your left") or feature-based (i.e., modifying information such as "those green and yellow ones"). As shown in figure 6, the mobile pairs produced a higher proportion of feature markings than the seated pairs. The odds of seated pairs producing feature over locative markings decreased by 63.5% (z = -2.10, p < .05). These findings support H4 and H5.

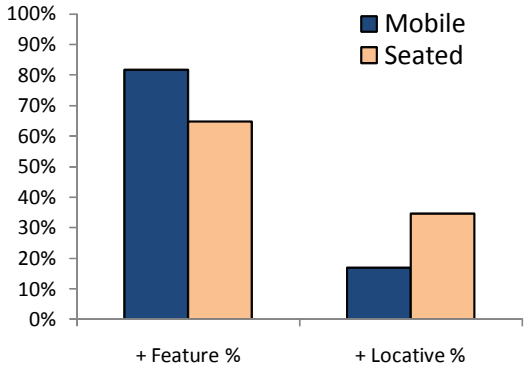


Figure 6. A greater portion of feature information was added by the mobile pairs compared to the seated pairs.

Gaze Coordination and Referential Form

We examined gaze patterns from 25 of the 33 pairs (7 side-by-side, 10 across, 8 mobile). The remaining pairs were withheld from analysis due to technical problems with recoding gaze or audio. The baseline measurements for individual fixations upon each of the sculpture objects were used to calculate a baseline gaze overlap rate of 8.28% for any given object.

The pairs demonstrated a significant degree of coordination in their gaze patterns during the reference phase. Both conditions exhibited a greater degree of gaze overlap than would be expected by chance (15.78% vs. 8.28%, t(23) = 8.45, p < .001) given the participants' overall distribution of gaze to the objects. Consistent with H6, the pairs demonstrated gaze coordination on the sculpture objects.

There were no detectable differences in gaze overlap across conditions ($F_{(1,25)}$ = 1.00, p = .33), and H7 was not

[2] Results report each variable's effect holding constant all other variables.

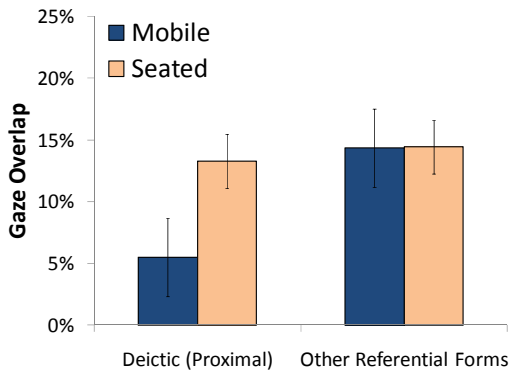


Figure 7. Gaze overlap (chance baseline is ~8.3%) by mobility condition and referential form.

supported. However, both remote ($F_{(1,1303)}$ = 18.26, p < .001) and local deixis forms ($F_{(1,1303)}$ = 10.03, p < .01) were correlated with lower rates of mutual gaze than other referential forms (e.g., definite and pronoun). A higher-order interaction reveals that this difference is even stronger for the pairs in the mobile condition when using local deictic references ($F_{(1,1302)}$ = 5.98, p = .015; see figure 7 for the interaction). Together these results offer partial support for H9 and H10.

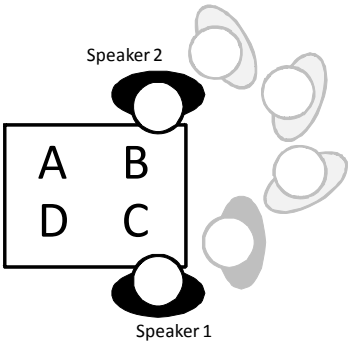
Additionally, it is worth noting that adding the referent shift variable to this model shows a significant effect of reference continuations on gaze overlap. When introducing a referent, gaze overlap is significantly lower than during subsequent mentions of that referent ($F_{(1,1271)}$ = 31.51, p < .001). Although this effect edges out the significant effect of local deixis, the local × seated interaction remains significant ($F_{(1,1276)}$ = 5.91, p = .015). These results offer partial support for H11.

Deixis, Addressee Gaze, and Topic Shifts

Finally, in looking at the interaction between gaze and discourse factors, we found that the speakers shifted focus to a new sculpture using deictic references when their addressee was gazing less at the intended sculpture beforehand (z = -2.32, p = .021). Speakers appear to monitor their addressees to determine if they need to coordinate attention. When their addressees are not attending to a referent, they may use referential expressions such as deictic demonstratives to help direct the addressee's attention. This provides support for H8.

DISCUSSION

We have mapped some of the ways in which reference, gaze coordination, and spatial context interact when people engage in collocated conversation. Free-standing, mobile pairs used more local deixis to refer to objects and had lower gaze overlap as compared to seated, stationary pairs. This reflects the mobile participants' access to movement as a coordination mechanism. Figure 8 demonstrates this during an exchange between mobile participants.



Speaker 1 finishes discussing object C.

Speaker 2 then begins to move away and walks to the other side of the table.

Speaker 2 then positions herself in front of Object B and the following exchange occurs…

Referring Expression	*Referent*	*Form*	*Ref. Shift*	*Gaze Overlap*
Speaker 2: **This one** seems like **it**'s very simply made too.	B	Local deixis	Init.	45.12%
Speaker 1: **It** does.	B	Pronoun	Cont.	96.59%
Speaker 2: There's just like a very basic structure to **it**.	B	Pronoun	Cont.	97.7%
Speaker 1: **It** feels like something I would have built using Legos.	B	Pronoun	Cont.	97.73%

Figure 8. Spatial evocation of a referent.

This excerpt demonstrates how the pair used a shift in positioning to evoke a new referent, allowing the speaker to shift the discourse focus using a referential form which, by itself, might not be sufficient for the speaker and addressee to coordinate their attention. From there, they are able to easily establish the "simply made" sculpture B as their current focus, and from there shift to a higher degree of gaze coordination and a reduced referential form ("it").

However, sometimes speakers needed to use additional linguistic information to evoke referents. When mobile pairs were using deictic references with added information, they were more likely to use feature-based descriptions, while seated pairs relied more on location-based descriptions. This suggests that mobile pairs use movement to coordinate attention to referents' location in space, but may be less inclined to use spatial descriptions due to the fact that their referential domain shifts. When the meaning of terms like "on the right" are variable, mobile pairs are more likely to use information about objects' appearance to direct attention when non-verbal coordination fails. Seated pairs, on the other hand, have a stable frame of reference, which allows for their increase in the use of locative terms to coordinate attention.

We also found that speakers used their addressee's gaze as an indication of attention. Speakers were more likely to use demonstrative deixis when they shifted the discourse focus to a new object and addressees were not looking at the speaker's intended referent. This highlights how speakers flexibly use different conversational resources to direct attention: when pairs have shared visual evidence, they rely

less on language to communicate, but when they are not coordinated in their visual attention, specific referential forms can be used to direct attention.

Design Implications

How can a multimodal understanding of reference be used to improve collaborative systems? Patterns of gaze and reference offer clues into how speakers perceive their shared visual space as well as their addressee's attentional state. If speakers use simpler referential forms, it suggests that their addressees have visual or spatial evidence that will allow them to disambiguate between all of the objects that a vague reference like "this one" could potentially refer to – perhaps prototypes on the conference table in a design meeting or graphical elements in a collaboratively edited document. Collaborative systems could better assess which objects in the environment users are talking about by weighing evidence about the likely referent of a referring expression. The addition of disambiguating information to deictic references suggests that a pair's attention may be divided between several similar objects in their shared environment, or that the speaker is prioritizing accuracy over efficiency in their communication, both of which could be useful in assessing a pair's interactions and task status.

Systematic differences in the use of deictic expressions (and larger differences in referential language) between spatial contexts could be very helpful in allowing collaborative systems to make inferences about the intent and attention of speakers. For example, automated camera systems for video-mediated collaboration [29] might combine information on the interlocutors' position and gaze patterns with the speaker's referential form to focus the camera on what is likely a speaker's intended referent. Probabilistic models based on findings like the ones in this paper (e.g., mobile users' tendency to say "this" when bringing up a new referent) could be used to assess topical shifts or attention shifts. Techniques such as these could help intelligent user interfaces to infer discourse focus, in turn helping systems to better track the status of a task or provide context-appropriate information.

Importantly, this study demonstrates the need for designers and builders of collaborative systems to acknowledge the systematic differences in collaborative reference between mobile and stationary users. Intelligent user interfaces using natural language processing systems designed for desktop contexts will not work as well in mobile contexts, and vice versa. We suggest that more research into the intersection of verbal and non-verbal patterns of language use is needed to design systems that understand human interaction in dynamic spatial contexts.

Limitations

We encountered several technical challenges with this new methodology such as the compounded problem of data loss when relying on two streams of gaze data. Our approach also requires some post-processing to ensure that data from multiple sources are temporally aligned and that gaze overlap comparisons are not being made between existing data points and dropped data points. We also encountered analytical challenges, such as modeling dyadic time-series data for pairs with different coordination styles and elaborating new parameters such as gaze overlap in the context of conversational initiative. Current approaches to quantizing and clustering gaze data are not well understood.

While we believe this work is an important step towards providing collaborative systems a human-like understanding of situated reference, several extensions to this work are needed. It will be important to further delineate the role of gesture as a part of referential communication in relation to gaze coordination and mobility. Subsequent studies should also vary additional facets of the groups and environment, as recent work by Bard et al. [4] suggests. Group composition characteristics (e.g., expertise or gender), object characteristics (e.g., lexical complexity), and environment characteristics, all are likely to play a role in referential form. Finally, with a deeper understanding of natural, multimodal reference, we will need to develop and test probabilistic models of reference for use in collaborative systems. Finally, we plan to implement and study the use of real-time dyadic eye tracking to harness gaze coordination as a form of input in collaborative, mobile systems.

CONCLUSION

In this paper we have provided several key contributions. We introduced a novel dyadic eye tracking methodology and a set of metrics for studying the multimodal process of reference. We also helped to advance theory on collaborative reference by identifying systematic differences between how non-verbal coordination mechanisms, and specifically gaze coordination and spatial context, affect referential language use in mobile and seated pairs. Finally, we described applications for our findings in the design of intelligent user interfaces for collaborative systems.

ACKNOWLEDGMENTS

We would like to thank Isaac Wilson, Steve Howard, Emilee Rader, Kathleen Geraghty and Patti Bao for their contributions. Funding was provided by National Science Foundation grants #0705901 and #0953943, and an NSF GRFP award to the second author.

REFERENCES

1. Argyle, M., & Cook, M. (1976). *Gaze and mutual gaze*. Cambridge: Cambridge University Press.
2. Ariel, M. (1988). Referring and accessibility. *Journal of Linguistics, 24*, 65-87.
3. Bangerter, A. (2004). Using pointing and describing to achieve joint focus of attention in dialogue. *Psychological Science, 15(6)*, 415-419.

4. Bard, E. G., Hill, R., & Arai, M. (2009). Referring and gaze alignment: Accessibility is alive and well in situated dialogue. *Proc. of the 31st Annual Conference of the Cognitive Science Society*.
5. Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory & Language, 59*, 457-474.
6. Boker, S. M., Xu, M., Rotondo, J. L., & King, K. (2002). Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. *Psych. Methods, 7(3)*, 338-355.
7. Bulling, A., Roggen, D., & Troster, G. (2008). It's in your eyes: Towards context-awareness and mobile HCI using wearable EOG goggles. *Proc. of UbiComp '08*, 84-93.
8. Byron, D. K., Mampilly, T., Sharma, V., & Xu, T. (2005). Utilizing visual attention for cross-modal coreference interpretation. *Proc of CONTEXT-05*.
9. Carlson, G. (2004). Reference. In L. R. Horn & G. Ward (Eds.), *The Handbook of Pragmatics* (pp. 74-96): Blackwell Publishing Ltd.
10. Carroll, J. M., Rosson, M. B., Convertino, G., & Ganoe, C. H. (2006). Awareness and teamwork in computer-supported collaborations. *Interacting with Computers, 18*, 21-46.
11. Cassell, J. (2004). Towards a model of technology and literacy development: Story listening systems. *Journal of Applied Developmental Psychology, 25(1)*, 75-105.
12. Chai, J. Y., Prasov, Z., Blaim, J., & Jin, R. (2005). Linguistic theories in efficient multimodal reference resolution: An empirical investigation. *Proc. of IUI '05*, 43-50.
13. Cherubini, M., & Dillenbourg, P. (2007). The effects of explicit reference in distance problem solving over shared maps. *Proc. of GROUP 2007*, 331-340.
14. Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22(1)*, 1-39.
15. Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology, 6*, 84-107.
16. Devault, D., Kariaeva, N., Kothari, A., Oved, I., & Stone, M. (2005). An information-state approach to collaborative reference. *Proc. of ACL 2005*.
17. Fussell, S. R., Setlock, L. D., & Parker, E. M. (2003). Where do helpers look? Gaze targets during collaborative physical tasks. *Proc. of CHI '2003 (Extended Abstracts)*, 768-769.
18. Gergle, D., Kraut, R. E., & Fussell, S. R. (2004). Action as language in a shared visual space. *Proc. of CSCW 2004*, 487-496. NY: ACM Press.
19. Gergle, D., Rosé, C. P., & Kraut, R. E. (2007). Modeling the impact of shared visual information on collaborative reference. *Proc. of CHI 2007*, 1543-1552.
20. Gundel, J. K., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language, 69(2)*, 274-307.
21. Hindmarsh, J., & Heath, C. (2000). Embodied reference: A study of deixis in workplace interaction. *Journal of Pragmatics, 32*, 1855-1878.
22. Kehler, A., Martin, J., Cheyer, A., Julia, L., Hobbs, J. R., & Bear, J. (1998). On representing salience and reference in multimodal human-computer interaction. *Proc. of AAAI '98*, 33-39.
23. Land, M., Mennie, N., & Rusted, J. (1999). The Roles of Vision and Eye Movements in the Control of Activities of Daily Living. *Perception, 28(11)*, 1307-1432.
24. Mutlu, B., Forlizzi, J., & Hodgins, J. (2006). A storytelling robot: Modeling and evaluation of human-like gaze behavior. *Proc. of Humanoids '06*.
25. Nakano, Y. I., Reinstein, G., Stocky, T., & Cassell, J. (2003). Towards a model of face-to-face grounding. *Proc. of ACL 2003*, 553-561.
26. Ou, J., Oh, K., Yang, J., & Fussell, S. R. (2005). Effects of task properties, partner actions, and message content on eye gaze patterns in a collaborative task. *Proc. of CHI 2005*, 231-240. ACM Press.
27. Pelz, J. B., Canosa, R., & Babcock, J. (2000). Extended tasks elicit complex eye movement patterns. *Proc. of ETRA 2000*, 37-43.
28. Prasov, Z., & Chai, J. Y. (2008). What's in a gaze? The role of eye-gaze in reference resolution in multimodal conversational interfaces. *Proc. of IUI '08*, 20-29.
29. Ranjan, A., Birnholtz, J. P., & Balakrishnan, R. (2007). Dynamic shared visual spaces: Experimenting with automatic camera control in a remote repair task. *Proc. of CHI 2007*, 1177-1186. NY: ACM Press.
30. Richardson, D. C., & Dale, R. (2005). Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cog. Sci. 29(6)*, 1045-1060.
31. Sidner, C., Lee, C., Kidd, C., & Lesh, N. (2004). Explorations in engagement for humans and robots. *Proc. of the Int'l Conference on Humanoid Robots*.
32. Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. E., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*, 1632-1634.
33. Torrey, C., Powers, A., Fussell, S. R., & Kiesler, S. (2007). Exploring adaptive dialogue based on a robot's awareness of human gaze and task progress. *Proc. of Human Robot Interaction*, 247-254.
34. Traum, D. R. (1999). Computational models of grounding in collaborative systems. *Proc. of AAAI Fall Symposium on Psych. Models of Comm.*, 124-131.
35. Ward, G., & Kehler, A. (2005). Syntactic form and discourse accessibility. In A. Branco, T. McEnery, R. Mitkov & J. Benjamins (Eds.), *Anapohora processing: Linguistic, cognitive and computational modeling*.