Tuning accessibility of referring expressions in situated dialogue

Short title: Tuning accessibility in dialogue

Ellen Gurman Bard	Robin Hill	Mary Ellen Foster	Manabu Arai
University of Edinburgh	University of Edinburgh	Technische Universität München	University of Edinburgh

Corresponding author: Dr E. G. Bard, Linguistics and English Language, School of Philosophy Psychology and Language Sciences, University of Edinburgh, Dugald Stewart Building, 3 Charles Street, Edinburgh EH8 9AD, U.K. Land line: +44 (0)31 651 1759. Email: ellen@ling.ed.ac.uk, Fax: +44 (0)131 651 3190.

This work was supported by the EU FP6 IST Cognitive Systems Program under Integrated Project "JAST: Joint Action for Science and Technology" (FP6-003747-IP). The University of Edinburgh and research sponsors are authorized to reproduce and distribute online copies for their purposes notwithstanding any copyright annotation hereon. The authors are grateful to Tim Taylor, Joseph Eddy, Jonathan Kilgour, Craig Nicol, and Jean Carletta for assistance with programming and to Jan Peter de Ruiter and Mark Steedman for helpful discussions. The faults remaining are entirely ours. Dr. Foster is now at the Department of Mathematical and Computing Sciences, Heriot Watt University, Scotland. Dr. Arai is now at the Department of Language and Computing Sciences, University of Tokyo, Japan.

Accessibility theory (Ariel, 1988, 1990, 2001) associates more complex referring expressions with less accessible referents. Felicitous referring expressions should reflect accessibility from the addressee's perspective, which may be difficult for speakers to assess incrementally. If mechanisms shared by perception and production help interlocutors align internal representations (Pickering & Garrod, 2007), then dyads with different roles and different things to say should profit less from alignment. We examined introductory mentions of on-screen shapes within a joint task (Carletta, et al., 2010) for effects of access to the addressee's attention, of players' actions, and of speakers' roles. Only speakers' actions affected form of referring expression. Only different-role dyads made egocentric use of actions hidden from listeners. Analysis of players' gaze around referring expressions confirmed this pattern: Only same-role dyads coordinated attention as accessibility theory predicts. The results are discussed within a model distributing collaborative effort under the constraints of joint tasks.

Key words: Referring expressions, accessibility, dialogue, joint action, cross-recurrent gaze

Tuning accessibility in dialogue

What a thing should be called (R. Brown, 1958) still engages anyone who deals with human or machine language production. One very wide-ranging approach (Ariel, 1988, 1990, 2001) relates the form of referring expressions to ease of access to the referent concept, discourse entity, object, state or event. Expressions introducing completely unfamiliar entities should be maximally detailed, as in, for example, indefinite NPs including modifiers (*a former Republican senator from strongly Democratic Massachusetts*). Expressions with more accessible referents might be definite NPs (*the red balloon)*, deictic expressions (*that cab*), or personal pronouns (*she*) in that order. Those referring to maximally accessible antecedents, usually the most recent in focus, can be as minimal as so-called clitics, unstressed and all-but-deleted pronouns (*/ts/ in the garage*). Accessibility theory offers a unified framework for predicting how forms of referring expressions will respond to givenness, discourse focus, and inferrability from discourse scenarios. In a similar theory, a Givenness Hierarchy of discourse phenomena defines referential form, with each additional degree of givenness for the referent permitting, though not demanding, a less elaborated expression (Gundel, Hedberg, & Zacharski, 1993, 2012). One way or another these models link the process of discovering referents with the form of referring expressions. To keep the members of this relationship distinct, we will treat accessibility as a property of referents and complexity as a property of referring expressions.

Although accessibility theory was developed to explain the relationship between anaphors and their antecedents (Ariel, 2001), many striking examples of complexity variation are found among expressions introducing a new entity into speech or text. Accounts of discourse phenomena are often framed in ways that make it possible to relate referring expressions to their real-world context, as well as their linguistic context (Prince, 1981).

Taken this broadly, accessibility ought to cover any conditions which might draw attention to the correct referent. To date, computational models of referring expression production largely target initial mentions of visible objects with competing possible referents (Dale & Reiter, 1995; Dale & Viethen, 2010; Krahmer & van Deemter, 2012; Vogels, Krahmer, & Maes, 2012) and generate the lexical contents of complex expressions with articles, adjectives and nouns. Work on human production of referring expressions has often dealt with anaphoric pronouns (Arnold, 2001, 2008; Fukumura & van Gompel, 2010; Fukumura & Van Gompel, 2012; Fukumura, van Gompel, & Pickering, 2010; Krahmer & van Deemter, 2012; Rosa & Arnold, 2011; Vogels, et al., 2012). While these starting points engender productive experimental paradigms and viable computational models, accessibility theory invites us to include in any model, psychological or computational, the forces that motivate choice among a wider range of forms. This paper examines the effects of referential accessibility on the form of first mentions in situated dialogue, where many forces are at play. We ask whose perspective constrains referential form, how speakers might track an addressee's referential perspective and whether, in situations where a speaker, an addressee, and context under discussion are co-present, there is any point in attempting to distinguish perspectives at all.

To be maximally useful, referring expressions should reflect the addressee's ability to discover the intended referent. As Clark and Marshall (1981) explained, assessing the addressee's true needs could require unmanageable effort. Since speaker and addressee may share a physical location, a discourse history, and any number of experiences, Clark and Marshall suggest that a speaker exploits such co-presence by treating his or her own knowledge of potential referents as a proxy for the addressee's. Ariel's (2001) notion of accessibility depends, in fact, on what the speaker supposes is the case, not on what is genuinely easier for the listener. Since speakers rely on estimates of shared knowledge

(Savitsky, Keysar, Epley, Carter, & Swanson, 2011), our first prediction might be that increasing indications of shared experience would increase the apparent accessibility of referents, and decrease the complexity of referring expressions.

Yet reliance on one's own knowledge as proxy for the listener's could be particularly risky for referring expressions. Though many forms of co-presence can remain constant throughout a conversation—for example, that the interlocutors are cousins baking a cake in Aberdeen— accessibility can change from one referring expression to the next. A speaker who adjusts her referring expressions (from *a speaker* to *her*, for example) on the basis of her own unfolding behaviour shows a capacity for rapid adjustment to changing referential accessibility. For referring expressions to be appropriately geared to an addressee, however, the speaker should be equally sensitive to the addressee's changing circumstances.

Genuine adaptation of production to addressees relates to phenomena of differing scope. Longer-term adaptations include the names that interlocutors jointly devise for innominate objects (Brennan & Clark, 1996; Metzing & Brennan, 2003) or pre-emption of a term by a particular referent (Kronmüller & Barr, 2007). Speakers can also respond appropriately to those medium-term characteristics of a referent, for example, the set of currently competing referents which a felicitous referring expression must exclude (van Deemter, Gatt, van Gompel, & Krahmer, 2012). Faster changing situations may find speakers failing to deploy listener models in a timely fashion (Bard, et al., 2000; Bard & Aylett, 2004; Barr & Keysar, 2002; Horton & Gerrig, 2002, 2005a, 2005b; Horton & Keysar, 1996).

Though swift responses to an addressee's situation are known in perception (Brown-Schmidt, Campana, & Tanenhaus, 2004; Hanna & Brennan, 2007; Hanna & Tanenhaus, 2004; Hanna, Tanenhaus, & Trueswell, 2003), proxy effects appear to be more robust than genuine adaptation to the listener's perspective in formulating referring expressions. For example, Bard et al. (2000) and Bard and Aylett (2004) found that speakers geared the phonetic counterpart of complexity, articulatory clarity (Lindblom, 1990), to their own knowledge, under a various modifications in the addressee's knowledge. In one study, speakers used shorter, less intelligible tokens of the same words when introducing to a new addressee (addressee-new) items which they had already discussed with someone else (speaker-old), while the elaboration of whole referring NPs, was more listener-sensitive. Galati and Brennan (2010) had speakers re-tell a story both to a new addressee (speaker-old addressee-new) and to the original addressee (speaker-old addressee-old). Word intelligibility in repeated narrations responded both to a proxy effect (with reductions for speaker-old addressee-new words) and to the identity of the addressee (with further reductions in speaker-old addressee-old tokens).Word duration did not differ by listeners' experience. Since several aspects of the addressee-old version were attenuated relative to the addressee-new version, it is not clear whether further degraded tokens in the addressee-old condition were keyed to accessibility or to violation of the maxim of quantity in a re-telling with no new information to convey.

In other studies, the complexity of referring expressions follows the speaker's local situation, not the listener's. Speakers use pronouns less often when there are more characters to distinguish, even if the referent could be uniquely distinguished by a gender-marked pronoun (Arnold & Griffin, 2007; Fukumura, et al., 2010). Fukumura and van Gompel (2012) have shown that changes in privileged knowledge can control relative incidence of pronouns and more complex expressions, while Rosa and Arnold (2011) found that speakers altered complexity of referring expressions in response to changes in their own ability to attend to referents, but not in response to their listeners' analogous distractions.

It is worth noting here that global adjustment to long-term co-presence— for example to the fact that the dyad are cousins—is not necessarily a series of locally tuned responses of the same type. Instead, local and global adjustments may differ in polarity or type, with the former more vulnerable to disruption by cognitive burden. For example, when another individual's attention is manifested in a visible eyetrack throughout a shared task, speakers' task strategies differ globally from those used in sessions without this cue (Bard, et al., 2007; Brennan, Chen, Dickinson, Neider, & Zelinsky, 2008). In a route communication task, however, Bard et al (2007) also found that speakers' use of the interlocutor's eyetrack was limited: they recognized that its arrival at the goal of an current instruction signalled its success, but were insensitive to its location elsewhere when it indicated an error, and they failed to make genuinely contingent responses. In a similar task , interlocutors made a behavioural distinction between global availability and local use of one another's faces (Anderson, Bard, Sotillo, Newlands, & Doherty-Sneddon, 1997). Overall, first mentions were less intelligible in dialogues with the interlocutor's face visible than dialogues with sight-lines blocked by a flimsy barrier. On the rare occasions when interlocutors actually looked at one another's faces, however, intelligibility increased to levels found when sight lines were blocked. In both cases local changes are constrained by some kind of cognitive burden. The distant eyetrack was ignored. The listener's face was checked only when the speaker had uncertainties to resolve (Doherty-Sneddon, Bruce, Bonner, Longbotham, & Doyle, 2002). Thus speakers can globally adjust forms of referring expression to a communicative situation where they might expect more information about their interlocutor's knowledge, while making a different adjustment locally when they need to use the information.

In sum, though speakers can adjust to longer-term co-presence or select forms of referring expression in rapidly changing situations from their own perspective, these capacities may not be regularly deployable at granularities needed to adjust the forms of referring expressions to their referents' accessibility from the addressee's perspective as a dialogue unfolds. These facts suggest that local adjustments would more often be based on the speaker's situation than the listener's.

What mechanism can register a speaker's own version of accessibility incrementally and still provide such a variable account of the listener's needs? Incremental operation is plausible under Expectancy (Arnold, 2001, 2008), the proposal that accessibility amounts to expectation that an entity will be mentioned, with complex referring expressions reflecting a kind of surprisal. Under this proposal reference production resembles syntactic processing (Demberg & Keller, 2008; Levy, 2008; Levy, Fedorenko, Breen, & Gibson, 2012) in depending on a probabilistic predictive mechanism. Both syntactic and discourse structure (Grosz, Joshi, & Weinstein, 1995) could affect expectations. The fact that narrative protagonists are often sentence subjects could be responsible for Fukumura and van Gompel's (2010) finding that speakers use a greater proportion of pronoun anaphors in referring to subject characters than in referring to object characters, though they themselves referred equally often to both. The difficulty here, however, is explaining how speaker and listener develop different views of a referent's accessibility. Statistically based expectations are usually the product of long-term experience of language and amount to a referential version of linguistic co-presence. Barring very different prior linguistic experience or simple inattention, it would take some other factor, like a different appreciation of the goals of the dialogue, acute sensitivity to different fast-changing local conditions, or statistical error in prediction to misalign expectations and create situations where a referring expression is not tailored to its addressee.

Pickering and Garrod (2004, 2007) make a suggestion that could lead to a more variable form of prediction. They propose that we deploy the same production processes in language comprehension and in language production. If our production system is recruited for prediction in comprehending language, and changes its state as it is used, it becomes the carrier of structural and lexical priming in dialogue. Interlocutors would start the planning of each new utterance with their production mechanism not in a neutral state but with the structures and lexical contents of the previous turn primed. The state of a speaker's production system in dialogue is therefore not identical to the addressee's, but somewhere between the speaker's own state in isolation and the current state of the addressee's production system. Whatever else we do to model our interlocutors' knowledge, it is difficult to imagine that an automatic, cost-free system like priming would not be recruited to lighten the potential cognitive burden of language production (Garrod & Pickering, 2004). With this baseline, a speaker might not only come to share the addressee's predictions, but also devote more active attention to the assessing addressee-specific information.~~n.~~

While this model would explain the advantage of interactive settings in sensitizing interlocutors to one another's perspectives (Brown-Schmidt, 2009; Ferreira & Hudson, 2011), it also provides a potential reason for some failures. Its utility would be curtailed when successive conversational turns do not draw on the same structures. In some settings, what interlocutors produce as prediction during comprehension could be quite dissimilar to what they need to produce aloud in conducting a dialogue. For example, the tangram matching paradigm typically assigns different tasks to describers and matchers. The describer has to provide referents with descriptions, producing referring expressions that evoke a distinctive visual analogy to the current innominate target tangram. The matcher has to provide descriptions with referents, confirming, querying, or negotiating details of description (Clark & Marshall, 1981). In contrast, the syntactic priming paradigm makes both participants describers and matchers for similar displays, so that they alternate in making descriptive statements of limited types. This is the paradigm that yields strong syntactic alignment. If Pickering and Garrod (2007) are correct in proposing a mechanism which should work best in even-handed situations, then interlocutors with different tasks will profit less in production from the benefits of perception.

Because the effects of structural priming are not long-lasting (Pietsch, Buch, Kopp, & de Ruiter, 2012; Potter & Lombardi, 1998; M. Smith & Wheeldon, 2001), a predictive-productive model will provide closely coupled but transient similarity between interlocutors. There is increasing evidence that a long-recognized tendency towards imitation during social interaction (Chartrand & Bargh, 1999; Meltzoff & Moore, 1977) involves close temporal coupling (M. Richardson, Marsh, & Schmidt, 2005; Shockley, Santana, & Fowler, 2003) of a dyad's activity. Close coupling in interlocutors' behaviour was observed in a third dialogue task, the map task, in which an instruction giver helps an instruction follower to reproduce a route pre-printed on the giver's map. In a face-to-face version of this task, players inadvertently imitated one another in linguistic, paralinguistic, and non-linguistic behaviours, at lags short enough to belong to a pair of adjacent dialogue turns (Louwerse, Dale, Bard, & Jeuniaux, 2012).

In summary, then, it is possible that natural priming in dialogue relieves speakers of some of the burden of tracking fine changes in their listeners' perspectives, with greater similarity in interlocutors' conversational roles permitting better keying of referential form to referents' accessibility to listeners. The effect might be carried by perception-production priming directly, or it might operate indirectly, with priming allowing more cognitively expensive processes to track the listener's perspective as the speaker formulates a referring expression. In either case the prediction is that similarity in production requirements should yield better keying of referential form to addressees' perspectives.

So far our questions about accessibility have largely ignored the setting in which referring expressions are used. A shared setting and, more important, shared attention will make proxy settings more useful approximations of addressees' needs. Observing that dyads viewing the same film often made initial mentions in very simple forms, Smith, Noder, Andrews, and Jucker (2005) concluded that shared experience of the referents could make them salient for speaker and listener, and keep them permanently in focus. Prominently moving or changing referents in a series of pictures also seems to provide focus, eliciting high accessibility pronominal forms (Vogels, et al., 2012) In such cases, speaking about shared visual experiences might provide few opportunities to refer to a genuinely inaccessible referent. There might be a range of forms in use, but in the Givenness Hierarchy (Gundel, et al., 1993, 2012) approach to accessibility, maximally accessible referents allow the full range of complexity in referring expressions. Any form could be suitable, because references to salient entities are unlikely to fail. In these circumstances, under-specification of a referring expression is not problematic.

These considerations suggest a number of forces which might affect the form of first mentions. Speakers might make first mentions simpler when indications of addressees' attention are available. They might respond incrementally to those indications when they actually occur. They might use their own situations as proxy for the addressees' if adjustments have to made locally. Or speakers might use proxy settings locally when they cannot benefit from automatic priming in production. Finally, speakers might share perspectives so closely that calculations of an addressee's perspective would be pointless.

To test for these possibilities, we use the Joint Construction Task or JCT (Carletta, et al., 2010), a task that is not inherently asymmetrical, or inherently an exercise in referring expression production, and that engages interlocutors in joint action within a common setting. Developed to study human-human joint action as a model for human-robot cooperation in quasi-industrial settings, the task requires two players to construct a two-dimensional tangram on their yoked computer screens (Figure 1). Each player can manipulate the components, all coloured geometric shapes, by mouse actions. Carefully coordinated action is critical. For two parts to be joined together, a different player must be moving each one. Parts join permanently where they first touch, so that careless alignment of edges can be remedied only by discarding the construction and starting again. If both players select the same object, or if two objects overlap, the component parts break and must be replaced, increasing the cost of the trial in both time and materiel.

(INSERT FIGURE 1 ABOUT HERE)

Figure 1 shows the JCT display early in a trial, when players have not yet begun to reproduce the target tangram (top right corner) from the movable parts (below the target tangram). These, the counters for time and breakages (top centre and left), the stock of spare parts (bottom), and the work space that occupies most of the screen, are always shared, with identical representations on the two players' screens. The goal is always reproducing the target tangram as accurately and economically in time and replacement parts as possible. No restrictions are placed on what players say. The forms of referring expression they use are under their control.

We examine the distribution of initial mentions of on-screen objects across the forms of referring expressions listed in Table 1, ranging from the most complex indefinite noun phrases, through definites, and deictics, to pronouns. Since all but one of the parts come in identical pairs (see the set beneath the target Figure 1), reference to either requires only type-identifiability, a minimum level of accessibility met by indefinite referring expressions (*a pink square*) (Gundel, et al., 2012). Accessibility theory predicts that decreasing complexity of form should accompany conditions that increase accessibility of referents.

(INSERT TABLE 1 ABOUT HERE)

To find global effects of access to an interlocutor's knowledge, we vary what each player can know about the other's attention. Players are both eye-tracked. If each player's eyetrack is cross-projected to the other's screen, each can locate the other's direction of gaze more precisely than in most real-world situations (Lobmaier, Fischer, & Schwaninger, 2006). Mouse locations can also be cross-projected. With either addition, a player can discover what his or her partner is attending to without waiting for the partner to move a piece or describe an action (Meyer, van der Meulen, & Brooks, 2004). Figure 1, for example, shows the partner's eyetrack (a blue circle) on the component parts where gaze must travel before a part is selected for removal to the main workspace. If referential form is adjusted to global accessibility, then providing gaze and mouse tracks should yield a result analogous to the findings of Anderson et al. (1997): references to on-screen objects should be in less complex forms in trials where some indication of addressee attention is cross-projected (Show Mouse or Show Gaze) than in trials lacking such indications (No Mouse or No Gaze).

To find effects of changing local conditions, we compare the forms of initial mentions which cooccur with speaker and addressee actions. Rather than an overall shift to lower complexity referring expressions, we expect a particular relationship between action and deictic forms.

When not anaphoric or contrastive, deictics typically require a demonstrator, a physical indicator of which part *this guy* or *that pink square* is (Clark, Schreuder, & Buttrick, 1983; Fillmore, 1982; Lyons, 1977). In the Joint Construction Task, a speaker can draw attention to an object by moving it or indicating it with the mouse while mentioning it. Foster et al (2008) showed that moving a part in this task acts as demonstration, promoting deictics over other forms. Because each player's mouse is a different colour and both change colour (compare Figures 2a and 2c) on selecting an object to move, players can distinguish whose mouse cursor is merely superimposed over a part and who has grasped it for movement.

Thus the paradigm allows for at least two kinds of demonstration: moving the referent and superimposing the mouse over it without grasping it, as long, of course, as the 'hovering' mouse cursor is visible to the other player. Though we have no independent evidence that hovering is pointing, it seems to be the closest thing to pointing that the JCT task allows.

We compare the expressions produced by dyads assigned the different roles of task manager and assistant with others who are both given the same role as partners in reproducing the target. If players can attune their perspectives better in identical roles than they do in different roles, poorer adjustment to the addressee's perspective is expected from those performing different roles. If an additional cognitive burden falls on the task manager, who typically plans local strategies, then any disadvantage for these dyads should be found more in the manager's referring expressions than in the assistant's. If the issue is alignment itself, then both Different Role players will show poor tuning to addressees.

Figure 2 shows what the other player will see when a speaker moves a part or merely hovers his or her mouse over it. The cross-projected mouse cursor appears as a small square with two white 'eyes' and changes colour when the mouse is used to select and move a part. A moving part, visible regardless of mouse cross-projection (Figure 2a for the Show Mouse condition and 2b for No Mouse), always provides shared support for deictic reference. If the mouse hovers, that is, if it is superimposed over the part without selecting or displacing it, deictic expressions are still supported by shared demonstration when the mouse's location is visible on the interlocutor's screen (Figure 2c for the Show Mouse condition). When mouse location is not cross-projected (Figure 2d for the No Mouse condition), the location of the hovering mouse is privileged to the speaker: it gives the addressee no way to know which of the two squares *this square* is. Deictic referring expressions should be infelicitous in this case. They should occur only if speakers tune form of referring expression to their own actions rather than to demonstration available to their addressees.

(INSERT FIGURE 2 ABOUT HERE)

We can use these facts to test for a relationship between local changes in referent accessibilty and the distribution of initial referring expressions across the forms listed in Table 1. Demonstration should coincide with an increase in deictics at the expense of all other categories. Speakers keying their referring expressions to direct local evidence of the addressee's attention might make this change when the addressee moves, looks at, or indicates the named part, as long as these signs of attention are available (movement in all conditions, gaze in Show Gaze, hovering in Show Mouse). Speakers keying their referring expressions to those of their own actions which are shared with the addressee would make analogous shifts in referential form when moving an object (Figure 2a and 2c) or hovering the mouse over it if the mouse cursor is cross-projected (Figure 2b). On the other hand, speakers who key referential forms to their own perspectives will respond to their own privileged actions (hovering the mouse over the referent in the No Mouse condition, Figure 2d) as well as to shared events (Figures 2a-c). Where we might predict less adjustment to listeners' needs (in Different Role dyads or in task managers), we would expect more such infelicitous expressions.

Our first investigation examines the distribution of first mentions of on-screen parts across the referential forms listed in Table 1, under varied global and local conditions experienced by JCT players operating in same or different roles. It asks which of the sources of referential accessibility influence the choice of referential form. Our second investigation exploits players' eyetracks to determine whether referential accessibility is uniformly high in this task, leaving forms of referring expression with little work to do in directing players' attention to the right objects.

Accessibility and Referring Expressions

Method

Task. As illustrated in Figure 1, the Joint Construction Task or JCT (Carletta, et al., 2010) offers to two collaborating players a target tangram, geometric shapes for reproducing it, a work space, a counter for breakages, a clock measuring elapsed time (top centre), and a set of replacement parts. The players' task is always to construct a replica of the target tangram efficiently, maximizing speed and accuracy and minimizing breakages. An accuracy score, superimposed over the built tangram at the end of each trial, measures its overlap with the target tangram. The timer and breakeage counter are updated continuously.

Participants can manipulate an object by left-clicking it with the mouse and dragging it or by right-clicking and rotating it. Players' mouse cursors differ in colour (Figure 1) and each changes colour (compare Figures 2a and 2c) when it selects an object to move by clicking on it.

Any part or partially constructed tangram 'held' by both players will break and must be replaced from the spare parts store to complete the trial. Moving an object across another breaks both. Objects can be joined only if each is held by a different player. Objects join permanently wherever they first meet. Inadequate constructions can be purposely broken and rebuilt from spare parts, incurring a cost in both parts and time.

Apparatus. In the present study, each participant sat approximately 40cm from a separate CRT display in the same sound-attenuated room. Participants faced each other, but direct eye contact was blocked by the displays between them. Participants were eye-tracked monocularly via two SR-Research EyeLink II head-mounted eye-trackers. Head-worn microphones captured speech on individual channels. JCT software, which allows mouse control of on-screen parts, recorded positions and status of all on-screen objects. Continuous audio and video records captured a full account of locations and movements of individual parts, constructed objects, and cursors. Composite Camtasia videos recorded all movements and audio.

Participants, design and materials. Sixty-four Edinburgh University students, paid to participate, formed 32 same-sex dyads who had never met before. Four further dyads were discarded because of technical failures. Sixteen of the 32 dyads worked under the Same Roles condition and 16 under the Different Roles condition. The experiment was run in tandem with another which examined the role of speaking in joint action. Each dyad therefore participated in 8 experimental conditions produced by the factorial manipulation of three binary communication modalities: Speech (Speaking v Non-speaking), Gaze (Show Gaze, with each player's current eye-track cross-projected onto the other's screen twice within each 42 ms cycle, v No Gaze, without cross-projection), and Mouse (Show Mouse, with each mouse cursor cross-projected to the other player's screen, v No Mouse, without mouse cursor cross-projection). Participants could always see their own mouse cursor and the moving parts and constructed tangrams. Gaze and Mouse Cross Projecction conditions were pseudo-randomised following a Latin square. Either the first four or the last four conditions were Speaking conditions. Since we are dealing with referring expressions, only conditions where players spoke are analyzed here.

In the 16 Different Role dyads, one participant was randomly designated manager and the other assistant. The manager was instructed to maximize speed and accuracy while minimizing cost in replacement parts, and to signal the completion of each trial. The assistant was to help and to confirm the completion signal. The 16 Same Role dyads had the same working instructions but were assigned no roles. Either player could signal completion. In all cases, trials ended when one player declared the construction complete by pressing the spacebar and the other confirmed. After declaration and confirmation, an accuracy score reflecting similarity (% appropriately oriented overlap) between the built and the target tangrams appeared across the built tangram.

Each dyad reproduced a different target tangram on each trial, or 16 different target tangrams, 2 per condition, with 8 of these produced in speaking conditions. No tangram resembled a nameable object. All trials offered the same set of 13 parts, comprising 2 copies of each of 6 shape-colour combinations (squares or right-angle isosceles triangles differing in size and colour) and a single yellow parallelogram. Each target tangram used 11 of the 13 available parts. All dyads encountered 4 different initial layouts of these 13 parts, counterbalanced across experimental items. The extra pieces differed from trial to trial.

Results

Task measures. For trials with speech, task measures are reported here. Below are the significant results of by-participant ANOVAs for Roles Assigned (Same, Different) x Mouse Cross-Projection (No mouse, Show mouse) x Gaze Cross-Projection (No gaze, Show gaze) on each of the the performance measures collected automatically by the Joint Construction Task software: trial duration, trial accuracy (overlap between built and target tangrams), and trial cost (in broken parts).

Different Role and Same Role groups built equally accurate matches to the model tangrams (91.2% v 93.4% overlap with models, $F(1, 28) = 2.73$, *n.s.*, $MS_e = 35.69$, $\eta_p^2 = .089$) at the same cost in broken parts per trial (2.1 v 1.8, $F1 < 1$), but Different Role dyads took longer on average to complete each task (216.1s) than Same Role dyads (180.7s) ($F(1, 28) = 4.91$, $p = .035$, $MS_e = 9{,}384{,}000$, $\eta_p^2 = .149$).

Over both groups, Show-Mouse and No-Mouse conditions also produced tangrams of equal accuracy (92.6% v 92%: $F(1, 28) < 1$), though cross-projection of the mouse cursor made for shorter dialogues (Show Mouse =186.9s v No Mouse = 209.9s : $F(1, 28) = 9.34$, $p =$

.005, MS_e = 1580000, η_p^2 = .25) with marginally fewer breakages (1.8 v 2.2: $F(1, 28) = 3.87$, $p = .058$, MS_e = 1.145, η_p^2 = .11).

There were no significant effects of gaze cross-projection for any task measure (Accuracy: Show Gaze = 93.1% v No Gaze 91.5%: $F(1, 28) = 2.33$; Duration: Show Gaze =192.6s v No Gaze = 204.2s : $F(1, 28) = 1.19$; Breakages: Show Gaze =2.05 v No Gaze = 1.96, $F(1, 28) < 1$).

Coding referring expressions. Dialogues were transcribed in standard orthography with one ChannelTrans (2006) channel per speaker. Audio, video, and transcription channels were lodged in aligned XML format. With purpose built coding software (Carletta, et al., 2010) that allowed simultaneous access to all events, each referring expression was time-stamped for start and end points. Then reference to any on-screen object was coded with a referent identifier for that object. Working from the composite videos, coders could use any material within a dialogue to determine referents. All referring expressions were then tagged for all the categories in Table 1. This system modestly expands the version applied to an earlier corpus of task-related dialogues (Bard & Aylett, 2004) on the basis of Ariel's work (Ariel, 1988, 1990). Because the classifications are based on linguistic forms, disagreement between coders is negligible. To avoid empty cells in further analyses, accessibility categories were collapsed into the four levels noted on Table 1: 0 (indefinite and bare nominal NPs), 1 (definite NPs), 2 (deictics and possessive pronouns), 3 (other pronouns and clitics). The table gives the short titles for these groupings. Only the initial mention of any on-screen object in a dialogue was analyzed further.

Coding: mouse actions. Mouse actions were recorded by the experimental software in terms of player, screen location and button presses (left button = select and move, right button = select and rotate). The software also recorded the location of each on-screen object at 42ms intervals. Because movement of referents aligns to the prosody of the referring expression, and is interpreted appropriately (Jesse & Johnson, 2012), only overlapping combinations of movement and expression were assessed. If the location of a player's mouse coincided with the location of an object currently being referred to, and either button was depressed, the referring expression was automatically coded as Move. All other cases were coded as No Move. If locations of mouse and referent object coincided and neither mouse button was depressed, so that the mouse was superimposed over the referent object without the power to move it, the expression was coded as Hover. All other cases were coded as No Hover.

Statistical method. As a uniform test for our predictions, we used multinomial logistic regression (hereafter MnLogR) to examine the effects of global and local variables on the distribution of initial mentions across the categories of referring expressions, as set out in Table 1 above. The statistic measures the capacity of each predictor, here all binary variables, to alter the distribution. More precisely, it measures any change in the ratio of the odds of each other category to the odds of a base category that coincides with a change in value of a predictor. The complexity category corresponding to minimal accessibility for referring to one of two paired tangram parts, indefinite expressions, was used as the baseline. MnLogR provides measures of coverage overall (χ^2 and Nagelkerke R^2) and estimates relationships by outcome category (*B, Exp(B), the odds ratio* and the Wald statistic), but does not assess random effects.

To avoid effects of small or empty cells in full factorial comparisons, we regularly ran separate regressions on subsets of the data that were predicted to behave differently (for example, the two levels of mouse cross-projection). Main effects and critical interactions were forced into the model, with backwards elimination of interactions that did not add to its coverage. Significant results are displayed graphically as proportions and are cited as percentages the text. Additional graphs present the measures that MnLogR actually works with, the log of the ratio of the odds of each other category to the odds of the base category (Indefinites). Variables are represented in tables in their baseline form, e.g. No Mouse or No Speaker Move. Wald tests of effects on individual forms of referring expression all have *df* = 1.

Global effects. The first analysis assesses the global effects of the independent variables on the distribution across referring expression categories (Indefinites, Definites, Deictics, and Pronouns) of all 1775 first mentions identified in the corpus. A full factorial design for Roles Assigned (Same Roles, Different Roles) x Mouse Cross-Projection (No Mouse, Show Mouse) x Gaze Cross-Projection (No Gaze, Show Gaze) x Player (A, B) was examined with backwards elimination (p to reject = .1, p to enter = .05) for Player (in Different Roles, B = manager, A = assistant) and all interactions. Player and interactions were eliminated. An equation based on the remaining main effects accounted for significant changes in the distribution of referring expressions (Likelihood ratio $\chi^2 = 67.54$, $df = 9$, $p < .001$, Nagelkerke $R^2 = .04$). If additional information about an interlocutor's attention simply increases expected accessibility of referents, we would expect a general shift from the more complex low accessibility forms toward the less complex high accessibility forms. If indications of attention are principally a medium for demonstrating the referents of deictic expressions, there should be a shift towards deictics but not towards less complex forms. . As Table 2 shows, the latter is observed. While gaze cross-projection accompanied no significant change in any category, mouse cross-projection altered the relative proportions of the two forms of intermediate accessibility, decreasing definite first mentions (45.2% of the 939 No Mouse expressions v 28.8% of the 836 Show Mouse expressions, *Exp(B)* = 1.78, Wald = 15.76, $p < .001$) and, to a lesser extent, increasing deictic expressions (No Mouse 28.9% v Show Mouse 42.4%, *Exp(B)* = 0.78, $p = .09$) relative to the most complex baseline indefinites (No Mouse 14.5% v Show Mouse 16.7%). Pronoun rates were unaffected (No

Mouse 11.3% v Show Mouse 12%). Thus when their mouse movements were visible to their partners, players effectively traded definite expressions for deictics, without using more pronouns.

(INSERT TABLE 2 ABOUT HERE)

In the absence of interactions, Roles Assigned showed a similar pattern. Dyads who had been assigned different roles also used fewer definites (34.5% of the 947 Different Roles expressions v 40.9% of the 828 Same Role expressions, *Exp(B)* = 1.58, Wald = 9.64, p = .002) in their initial mentions and more deictics (Different Roles 35.9% v Same Role 34.7%, *Exp(B)* = 1.33, Wald = 3.67, p = .055) relative to baseline indefinites (Different Roles 17.8% v Same Role 13%). Again the rate of pronouns did not change (11.4% v 11.8%).

***Local effects*: *Speaker actions*.** Felicitous use of deixis depends on whether listener and speaker can both identify the referent via a demonstrator in the form of movement of the referent or of a visible gesture. We report separate analyses for conditions with mouse cross-projection (Show Mouse, n = 836) and without (No Mouse, n = 939) because they provide different access to demonstrators. The predictors included the experimental variable Roles Assigned, the speaker's mouse actions (moving the part being mentioned, hovering the mouse over it without selecting it), and the interactions of Roles Assigned with each movement variable. Gaze cross-projection was not included here, as it had proved ineffective in the global analysis and in earlier exploratory regressions. Interactions except for Speaker Hover x Roles Assigned could be eliminated if unhelpful to the model's coverage. Table 3a and 3b show the regression outcomes. Both models accounted for significant variation in distributions (Show Mouse: χ^2 = 45.635, df = 12, p < .001, Nagelkerke R^2 = .058; No Mouse: χ^2 = 81.622, df = 12, p < .001, Nagelkerke R^2 = .091).

(INSERT TABLE 3 ABOUT HERE)

As predicted, actions obvious to speaker and listener were important. First, players moving the referent (Figure 3) used more deictic expressions at the expense of indefinites than they used when not moving referent objects. Since movement can always be seen, a change in the rate of deictics is expected whether or not the speaker's mouse cursor itself was visible, (Figure 3a, Show Mouse: 53% of the 301 referents being moved by the speaker and 37% of the 535 referents not being moved had deictic first mentions, Exp(*B*) = 0.49, Wald test = 8.67, *p* =.003; Figure 3b, No Mouse: 39% of the 611 being moved and 24% of the 328 not being moved, Exp(*B*) = 0.463, Wald test = 8.89, *p* = .003). Second, there is an effect of hovering a visible mouse cursor over the referent. In the Show Mouse condition (Figure 4a), the players used fewer pronouns in referring to objects under their mouse than in referring to others (6% of the 297 items with hovering v 15% of the 539 items without hovering, Exp(*B)* = 3.39, Wald test = 7.23, *p* = .007). Their tendency to use indefinites was unchanged (16% with v 17% without). The concomitant increase in deictics, (51% with hovering v 39% without), though on the same scale as with movement, did not reach significance (Exp(*B*) = .781, Wald test = 0.725). As Table 3a shows, there were no significant interactions: Same Role and Different Role dyads behaved alike (Roles Assigned x Speaker Move was eliminated, *p* > .05, Roles Assigned x Speaker Hover: Definites: Exp(*B*) = 0.708, Wald test = 0.553, Deictics: Exp(*B*) = 1.422, Wald test = 0.649, Pronouns: Exp(*B*) = 0.648, Wald test = 0.442.)

(INSERT FIGURE 3a and 3b HERE)

(INSERT FIGURE 4a and 4b HERE)

In contrast, actions known only to the speaker were important only for Different Role dyads. In the No Mouse condition (Figure 4b), when the speaker's mouse hovered over the referent, visibly to the speaker but invisibly to the listener, Different Role dyads all but abandoned indefinites (8% of the 192 referents introduced with hovering v 22% of the 287 introduced without hovering) for more deictics (40% with hovering v 25% without) (Speaker Hover: Exp(*B)* = .327, Wald test = 10.27, p = .001; Speaker Hover x Roles Assigned: Exp(*B)* = 3.32, Wald test = 6.74, p =.009) and more definites (48% with hovering v 36% without; Speaker Hover: Exp(*B)* = .304, Wald test = 13.19, $p < .001$; Speaker Hover x Roles Assigned: Exp(*B)* = 3.53, Wald test = 7.69, p = .006) than they used without this gesture. The numerical decrease in pronouns was not significant (4% with v 16% without hover ; Speaker Hover: Exp(*B)* = 1.55, Wald test = .797*;* Speaker Hover x Roles Assigned: Exp(*B)* = 2.23, Wald test = 1.46). As the leftmost panel of Figure 4b shows and the Speaker Hover x Roles Assigned interactions indicate, the Same Role dyads did not follow this pattern (indefinites 12% with v 12% without hovering; definites: 52% with v 48% definites without hovering; deictics: 31% with v 24% without).

To determine whether the behaviour of Different Role dyads was restricted to the manager, whose cognitive burdens might be greater than the assistant's, we added Player to the predictors in a MnLogR for Different Roles dyads in the No Mouse condition (Speaker Move x Speaker Hover x Player). The critical Player x Speaker Hover interaction was forced into the equation but other interactions could be eliminated as before. The resulting model, summarized in Table 4, provided significant coverage of the 479 first mentions made by Different Role dyads in the No Mouse condition ($\chi^2 = 52.61$, $df = 12$, $p < .001$, Nagelkerke R^2 = .113) but there were no effects of Player on the frequency of any referential form and, critically, no difference between between managers and assistants in the effects of an invisibly hovering mouse (Speaker Hover x Player: Definites, Exp(*B)* = .541, Wald test = 0.912 *n.s.*; Deictics, Exp(*B)* = .494, Wald test = 1.123). Effects of hovering on rates of definites and deictics are now Definites, Exp(*B)* = .439, Wald test = 3.647, p = .056*,* Deictics*,* Exp(*B)* = .462, Wald test = 2.888, p = .089).

(INSERT TABLE 4 ABOUT HERE)

Local effects*: *Addressee actions. Genuine effects of the listener's actions on the complexity of referring expression should appear in conditions where the speaker can notice those actions and where likely confounds are eliminated. None of the listener's actions—moving the referent, hovering the mouse over it, or looking at it —yielded effects meeting these criteria. Results are given here for the most complete appropriate models that could be calculated.

Cases where the addressee moved the referent should always have been noticeable. Because the JCT rules make it impossible for both players to move the same object and dangerous for both to reach for the same object simultaneously, empty cells prevent the development of a full model. The 891 expressions referring to items that the speaker neither moved nor hovered over could be modeled with Mouse Cross-Projection (2), Listener Move (2), and Roles Assigned (2) as predictors ($\chi^2 = 45.855$, $df = 21$, $p = .001$, Nagelkerke $R^2 = .054$), but the addressee's action did not affect the distribution of referring expressions (Definites; Exp(*B)* = 1.395, Wald test = 0.423; Deictics: Exp(*B)* = 1.099, Wald test = 0.038*; Pronouns:* Exp(*B)* = .957, Wald test = 0.006).

Cases where the addressee hovered the mouse over the referent without moving it should have been useful to the speaker only in the Show Mouse condition, when each player could see the other's mouse cursor. An overall analysis using Mouse Cross-Projection (2), Speaker Move (2), Speaker Hover (2), and Listener Hover (2) ($\chi^2 = 204.188$, $df = 45$, $p < .001$, Nagelkerke $R^2 = .118$) showed that the addressee's gesture was associated with increased likelihood that the speaker would use a pronoun (14% with v 11% without; Exp(*B)* = .192, Wald test = 3.90, $p = .048$). Yet there was no difference in this tendency between the Show Mouse condition,where the speaker could see this gesture and the No Mouse condition, where the s/he could not (Show Mouse x Listener Hover: Exp(*B)* = .1.23, Wald test = .019). Whatever encouraged speakers to use pronouns in these instances did not depend on their seeing the addressee's gesture and, in fact, had the opposite effect to other potential demonstrators: no other produces an increase in pronouns.

Cases where the addressee looked at the referent could have been observed by the speaker only in the Show Gaze condition, where eyetrack cursors were cross-projected. Because a player's gaze would have been attracted by a moving part or mouse, Listener Gaze, Speaker Move, Speaker Hover and Roles Assigned cannot be used as predictors of referential form in the same model. For the 891 referrring expressions performed without the speaker moving or hovering over the referent, speakers used more definite expressions to introduce referents that the addressee was looking at (51% with listener gaze) than otherwise (34% without listener gaze), (Predictors: Gaze Cross-Projection x Roles Assigned, and Listener Gaze, with Gaze Cross-Projection x Listener Gaze: $\chi^2 = 35.704$, $df = 12$, $p < .001$, Nagelkerke $R^2 = .042$; Definites Exp(*B)* = .214, Wald test = 7.794, $p = .005$). Again, however, the effects were the same whether the speaker could see the addressee's eyetrack cursor or not (Listener Gaze and Gaze Cross-Projection ; Exp(*B)* = 2.56, Wald test = 1.93, *n.s*). Whatever controlled the speakers' choice of expression here did not depend on their seeing where the addressee was looking.

Discussion

This study offered several ways to make referents accessible. First, they were on-screen, available to both players and critical to their joint task. Second, if players' gaze cursors were cross-projected, each could see where the other was looking. Third, if players' mouse cursors were cross-projected, each could see where the mouse was moving, pausing or grasping. Fourth, if either player moved an object, the other could tell that he or she was engaged with it under all conditions. In this setting, a speaker might frame referring expressions in accordance with global trial conditions or with local attention or action.

Sensitivity to the addressee might appear in two ways. There might be a shift towards the simpler forms of first mentions arising from an impression of greater accessibility when speakers had access to the listener's eyetrack or cursor. Or referring expressions could be keyed to local action, in particular to actions that could serve as demonstrations for deixis.

Of the cross-projection conditions, only the cross-projection of mouse cursors appeared to affect the form of first mentions of on-screen objects. The result was not an overall shift towards less elaborate forms, but a shift towards deixis: Definites were rarer, and deictics more common when mouse cursors were cross-projected.

Local effects show whose mouse was responsible. When no speaker actions coincided with addressee actions, form of first mentions was insensitive to what the speaker could see the addressee doing. Instead all players used fewer definite expressions or pronouns and more deictics when they themselves moved the referent. Finally players with different assigned roles were even less sensitive to the addressee's perspective: they also used more deictics when superimposing an unseen mouse over the referent. Thus, both of the shared demonstrators sketched in Figure 2a-c (moving a part and visibly hovering the mouse over it) affected all players' choices, while the privileged demonstrator (superimposing the mouse on the part when the mouse was not cross-projected, as in Figure 2d) affected those playing different roles.

Where Different Roles dyads produced infelicitous referring expressions, there was no distinction between manager and assistant as there would be if the designated manager had assumed a disruptive cognitive burden with the role. Instead, the use of deictic introductory mentions with privileged demonstration characterizes the dyad.

So far, the results suggest an association between role similarity and the ability to maintain felicitous use of refering expressions. But are the deictics with invisible demonstrators genuinely infelicitous in a rich task environment where speaker and listener should be coordinating their actions and attention carefully regardless of what they say to one another? Overall we did find a significant performance deficit for Different Role dyads, who took 20% longer on average than their Same Role counterparts to assemble the same tangrams to the same level of accuracy . To discover whether their difficulty had anything to do with using referring expressions to manage speakers'attention, our second investigation probes the relationship between referring expressions and players' attention.

Aligned gaze

Whatever the relationship between cognitive accessiblity of referents and formal complexity of referring expressions in texts, there is reason to suspect that accessibility might be less important in genuine joint action. With extended periods of joint physical action, accessibility for the objects of that action might stay equally high for both speaker and listener. If so, there is little disparity of knowledge to accommodate linguistically, and a speaker can safely use his or her view as a proxy for the addressee's. Two aspects of the JCT might be of concern here.

First, simply viewing the same arrays while conversing or hearing the same narration encourages dyads to coordinate their visual attention to some degree (D. Richardson & Dale, 2005; D. Richardson, Dale, & Kirkham, 2007). Second, the precise physical coordination required for the JCT leaves little to chance: dyads must attend to the same objects at the same time in order to move them into the correct positions and join them without breakages or mis-assembly. A speaker's gesture, private or public, might be irrelevant if players are working in a truly coordinated fashion.

In effect, the objects of coordinated joint action would be in discourse focus (S. Smith, Noda, Andrews, & Jucker, 2005) and highly accessible. In the view of Gundel et al (1993, 2012), more accessible referents licence not just less complex forms of referring expression, but a range of forms including the less complex. Extra-linguistic referents which seize attention have this character: the singing drunk attempting to walk in front of your moving vehicle can be as effectively denoted by *he* or *that guy* as by a colourful description of his intellect, musicality, and blood alcohol level.

Thus dialogue within joint action might be the wrong domain for calculating the importance of linguistic behaviour, distinguishing speaker and listener knowledge, or finding any orderly relationship between cognitive accessibility and form of referring expression. To determine whether this is the case here, we use gaze as an indicator of focussed attention.

If speakers' gaze tours objects about to be mentioned (Meyer, et al., 2004) and both players' gaze is largely and consistently focussed on the same objects when referring expressions begin, whatever form is generated, then the players' task will keep referent accessibility largely high and shared. If, instead, alignment in interlocutors' attention is not uniformly high, but differs across referring expressions, then accessibility differences may be available even in dialogue situated in a joint task. Finally, if differences in roles have the misaligning effects that we claim for them, we should see a less orderly relationship between form of expression and shared attention in Different Role dyads than in Same-Role dyads.

To assess joint attention, we measure alignment of players' gaze via the cross-recurrence analysis used by Richardson, Dale and Kirkham (2007) and Richardson and Dale (2005). Cross-recurrence (Zbilut, Giuliani, & Webber, 1998) measures both absolutely simultaneous activity and activity that may be entrained, that is, linked in type and timing, but not perfectly simultaneous. Figure 5 gives an example of such activity. The graph plots time from a single time signal for each of two individuals, one on each axis. Let us suppose that 0,0 is the beginning of musical accompaniment, and the diamond-shaped points labelled N, G on this graph show when two dancers in the Kirov Corps de Ballet, Natalya and Galina, complete each choreographed step. The two dancers should be exquisitely coordinated: we should hear only one clunk of blocked toe shoe on wood as two of them touch the floor. Thus, all their footfalls in any performance should lie along the diagonal which indicates simultaneous action. Now imagine that the square points labelled V, S show where Vicky and Sue, two friends who enjoy line-dancing at their local pub, complete their steps. Though they are doing the same steps to the same music, we expect to hear two clacks of cowgirl boots against the floor as they land. If Vicky lands first, the point V,S is to the left of the diagonal. If Vicky is copying Sue, however, then there will be many V,S points to right of the diagonal and fewer to the left. Cross-recurrence analysis plots the frequency of all N,G or V,S points with the diagonal in this graph becoming the 0 point in the cross-recurrence graph and the distance at right angles to this diagonal, the lag between actions, becoming the horizontal axis. Natalya and Galina should produce a distribution which peaks sharply at 0 lag and rarely attracts the ire of the ballet master by notable asynchrony. Under much less supervision and a few more beers, Vicky and Sue should should produce a flatter distribution, with a distinct asymmetry, a curve higher on the Sue-first side than on the the Vicky-first side, if Vicky is indeed copying Sue. For our own study we ask whether all forms of referring expression arise with visual coordination as sharp as in Natalya and Galina's case, or whether visual coordination depends on the nature of the dyad (as in Figure 5) and the kind of referring expression being produced.

If JCT dyads have high simultaneous gaze peaks whatever their roles and whatever forms of expression they use, then the task is providing joint focus and referring expressions have little work to do in directing players' attention. If coordination is less precise, then one player is seeking out an on-screen object before the other. If form of referring expression is well geared to accessibility from the listener's perspective, then order of players' gaze at referents should be related to the form of the referring expression. More complex expressions should be used largely to draw a listener's attention to objects which the speaker already has in view and which the listener must subsequently find. Less complex referring expressions should be found when the listener already has the referent in view.

INSERT FIGURE 5 ABOUT HERE

Method

This investigation uses the Joint Construction Task dialogues described earlier. It examines the coordination of players' gaze as they began to refer to objects on their yoked screens.

Cross-recurrent gaze. The regions of interest (ROI) for gaze were both fixed (the clock, penalty counter, target tangram, spare parts store) and dynamic (the movable parts and tangrams under construction). Fixations on blank areas of the background, looks off-screen and blinks were excluded. Each player's gaze location was examined for a temporal window of ±4s from the onset of a referring expression. To assure that gaze was associated with individual referring expressions, expressions were used only if the 4s prior to onset contained no other referring expression. This was the maximal gap we could allow for pre-utterance gaze at the ROI while still including the bulk of the assessed referring expressions. Analyses were based on 936 referring expressions. Table 5 shows how they were distributed across Roles Assigned and referring expression category.

(INSERT TABLE 5 ABOUT HERE)

Each player's sampled gaze was located at increments of 20ms before being pooled into bins of 200ms running from 4s before to 4s after the onset of the referring expression. With the speaker's gaze locations as a reference in time and space, the other player's gaze at each bin before and after the referring expression onset was checked for a match to the speaker's ROI. The likelihood of overlap between participants' eye movements was therefore examined when they lagged each other by up to four seconds.

Results

Figures 6a and 6b show the cross-recurrent gaze results. The *y*-axis shows percentage of fixations coinciding, while the *x*-axis shows lag from one player's fixation to the other's. At 0 lag, gaze is simultaneous. At negative lags, the listener's gaze reaches the ROI before the speaker's. At positive lags, the speaker's gaze arrives first. Curves are distinguished by form of referring expression. For each level of accessibility/complexity, filled points represent real cross-recurrent gaze, reflecting the time lag between interlocutors' matching fixations on any ROI, while corresponding unfilled points represent baseline cross-recurrent gaze for expressions of that referential form. The baseline was generated by randomly reordering one player's gaze records and running a cross-recurrence analysis with the other player's real record. This baseline reflects the probability that two individuals will look at the same objects purely by chance, given topic, form, or task. The randomly paired control is therefore typical of the situation in which the measurements were made, while the correctly ordered recurrence curves express both situation and timing. Significant differences between corresponding real and random curves would indicate temporal gaze coordination beyond chance.

(INSERT FIGURE 6 ABOUT HERE)

The first stage in testing the utility of accessibility in situated dialogue is to determine whether gaze coordination between interlocutors is always high. The second is to determine whether any discovered patterns in gaze coordination relate in an orderly way to referential behaviour, either in terms of speakers' tendencies to respect listeners' needs or in terms of referential form.

Figure 6 shows that gaze coordination was not always high. If each speaker were choosing randomly and independently among 15 referents (up to 13 movable parts and the store and target tangram), the chances of gaze overlap would be well under 1% as the trial begins and about 2.7% when 11 parts have been combined into a single new tangram. The randomized baseline curves (unfilled points on Figures 6a and 6b), which remove temporal coordination as a criterion for coinciding gaze, range from 28% to 33%, showing that the players were attracted to the same objects much more than chance would predict. These baselines differ because players can look at the same objects more or less in different conditions. The tendency to look at the same objects at about the same time (filled points on Figures 6a and 6b) peaks at the higher level of 30-43%, but still indicates that less than half of simultaneous fixations are directed to the same object.

Role effects. ANOVAs were run on real cross-recurrence values (filled points in Figure 6) and on real–random differences (filled – unfilled points) for individual referring expressions nested in Form of Referring Expression (4), Roles Assigned (2), Mouse Cross-Projection (2), and Gaze Cross-Projection (2) and crossed with Lag (41). Mouse and Gaze Cross Projection were included primarily to reduce error variance. Because of the similar ranges in randomized cross-recurrence (unfilled points), the same significant effects were found in both analyses. We report only the real–random results.

The question here is whether Same-Role and Different-Role dyads, who used referrring expressions differently, also align attention differently. Visual inspection of Figure 6 suggests that the groups differ in degree of aligned gaze and in its temporal symmetry. Maximum gaze overlap was 43% for Same Role dyads and 37% for Different Role dyads. As these peak values suggest, Same Role dyads achieved more gaze alignment averaged over the whole sampled period for (35.5%) than Different Role dyads did (31.3%) (role: $F(1, 904) = 12.83$, $p < .0005$, $MS_e = 3426$, $\eta_p^2 = 0.013$), as well as more pronounced alignment peaks of roughly simultaneous gaze (role x lag: $F(40, 36160) = 1.57$, $MS_e = 56.4$, $p < .02$, $MS_e = 56.4$, $\eta_p^2 = 0.002$). Thus, the group who used referring expressions more egocentrically aligned attention worse.

To determine whether Different Role dyads failed to coordinate because they simply did not look at the namable ROIs, we compared the groups' time looking off screen or at the clock. Different Role dyads did spend more time looking off screen ($F(1, 501) = 14.31$, $p < .001$) or at the clock ($F(1, 501) = 6.39$, $p = .012$) than Same Roles dyads, though not enough more (2-3% of fixations in total) to make coordinated on-screen gaze impossible.

Gaze coordination and referential behavior. As Figure 6 shows, Form of Referring Expression had no overall effect on gaze alignment ($F(3, 904) = 1.71$, $MS_e = 3246.2$) and no overall interaction with Roles Assigned ($F(3, 904) = 1.45$, $MS_e =3246.2$). There were, however, different temporal patterns of gaze alignment across Forms of Referring Expression that were dependent on Roles Assigned (Lag x Roles Assigned x Referring expression complexity: $F(120, 36160) = 2.60$, $p < .0001$, $MS_e = 56.4$, $\eta_p^2 = 0.009$). To understand these patterns we return to the basic principle of accessibility: that speakers should use more complex forms of expression for felicitous reference to entities that listeners will find harder to access *a priori*.

This principle predicts a trend in cross-recurrence: complexity of referring expression should predict the lag between the speaker's and the addressee's attention to the referent, with speakers attending before listeners whenever expressions need to be complex to direct listeners' attention to the correct inaccessible referent, and with listeners already attending to an in-focus accessible referent whenever referring expressions can be minimal. This trend would signal sensitivity to the addressee's needs. To look for the trend, we examined the tent-like structures of the real cross-recurrence curves for asymmetry. When one side of the peaked curve is higher than the other, there was more of the situation we described for Vicky copying Sue in Figure 5: one gaze was copying the other. To provide a measure of the degree of asymmetry in a real cross-recurrence curve, we subtracted speaker-first percentages of aligned gaze (shown at positive lags in Figure 6) from the corresponding listener-first percentages (at the negative lags), and averaged the 20 differences calculated this way for each curve. For example, the highest curve in Figure 6a, (filled triangles) represents real cross-recurrent gaze for pronouns. The left half of the curve, where listeners looked at the object before speakers did, is higher than the right half, where speakers looked first. Figure 7 shows the mean values for the four categories of referring expression in Same Role and Different Role dyads. A positive value in Figure 7 means that, as in the case of pronouns in Figure 6a, there were on average more instances when listeners looked first, while a negative value here means that on average there were more instances when speakers looked first.

(INSERT FIGURE 7 ABOUT HERE)

For the Same Role dyads, whose usage had respected shared information, the results follow the prediction: the more accessible a form's referents should be, the stronger the tendency for the addressee to look at them before the speaker (Spearman's ρ =.104, $N = 423$, 1-tailed $p = .016$). For the Different Role dyads, whose usage had been infelicitous, there was no orderly relationship between form and gaze alignment and there were no reliable differences between individual categories (Spearman's ρ = -.052, $N = 513$).

Discussion and conclusions

In this work, we attempted to discover what determines form of referring expressions in dialogue. We tested a simple prediction about co-presence: that having more information about a listener's attention should enhance apparent co-presence, increase apparent accessibility of referents, and elicit less complex forms of referring expression. This was not the global outcome we observed. We also tested for effects of local events on forms used for first mentions. There was no direct evidence that observable actions or attention on the listener's part attracted demonstrator-like effects, a shift towards deictic forms and away from personal pronouns. As in earlier experimentation in other paradigms (Fukumura & Van Gompel, 2012; Rosa & Arnold, 2011), there was evidence for adherence of referential form to the speaker's own perspective. Here the effect appeared to follow a prediction derived from involvement of productive mechanisms in perception (Pickering & Garrod, 2004, 2007): where perception and production of language employ the same production system, the benefits of alignment should be less for dyads whose production has to be disparate. Our findings supported this view: all dyads used more deictic expressions when using a demonstrating action that the addressee could observe, but only Different Role dyads provided clear examples of speaker-oriented design. They used more deixis when they hovered their mouse over the target than when they did not, even if the mouse was invisible to the listener. The result was not a matter of cognitive load, in so far as task managers, who should in general have had more responsibilities for planning strategy, were no more speaker-oriented than task assistants.

We then considered the possibility that apparently egocentric use of referential form might be irrelevant in richly situated dialogue where the context could supply what a dialogue history lacked. If this were the case, situated dialogue – dialogue produced in surroundings which interlocutors can see and interact with – would not be a useful domain for attempting to separate the perspectives of speaker and listener. At least for discussions of the here and now, co-presence would be more than a helpful shortcut to estimates of common ground; it could become an account of shared accessibility. In tasks like the JCT, which demand precise coordinated manual control of converging objects, patterns of attention might be so highly coordinated that the acted-upon world would be equally – and highly – accessible to the joint actors. If in-focus referents actually permit the full range of forms

(Gundel, et al., 1993, 2012), differences in form would not be attributable to cognitive demand, but to discourse rules or stylistic decisions on the pragmatic features to be emphasized.

Our findings suggest, however, that joint action does not coerce interlocutors' attention into a common pattern. Instead genuinely shared attention can vary both between and within dyads. Different Role dyads, who had accommodated accessibility to private gestures as well as public, did not coordinate attention well. Figure 6b shows that only definite references display the 'tent' shape peaking at simultaneous joint gaze, though the tendency did not reach significance. The order in which players looked at the referents formed no pattern. In contrast, the Same Role dyads produced visible cross-recurrence peaks for every referential category (Figure 6a) and their gaze sequence followed the order that accessibility would dictate if it were geared to the listener (Figure 7): the more accessible a referent should be to the listener if a particular form of referring expression is used, the more listeners' gaze actually preceded speakers'.

The results are consistent with a model of referring expression formation that permits a different time course for application of speaker and addressee information and that distributes effort and risk across a dyad. We propose that speakers assess both the likelihood of common ground in a dialogue overall and the likely risk to the dialogue's goals posed by misunderstanding. Greater likelihood of common ground makes speakers surer that they are communicating successfully, even if they are not (Savitsky, et al., 2011). Situated joint action should suggest copious common ground and invite proxy use. Whenever common ground fails in such situations, the dyad experience a sudden fall in the predictability of events, a phenomenon known as surprisal. Surprisal would disrupt the dialogue, but preventing common ground surprisal when producing speech could be more demanding still (Clark & Marshall, 1981). In the JCT, at the least, anticipating failures of common ground demands identifying what the addressee can see. A player using her own knowledge as proxy for common ground acts as if her addressee can always see her mouse because she herself can always see it. A player monitoring of the addressee's perspective must infer from absence: when she cannot see the other player's mouse, she must infer that the other player, therefore, cannot see hers.

Though the work required to take inferred facts into account continuously will be worthwhile in some kinds of dialogues—those with dangerous penalties for mis-communication, like emergency services calls, for example —the JCT risks little but additional time and shared effort. If a JCT speaker selects a form of initial referring expression which turns out to be underspecified for her addressee, and if that expression needs to be understood, then responsibility for making good the shortfall is readily shared with the addressee (Carletta & Mellish, 1996). The more information and attention the dyad actually share, the lower the chances of failed reference should be, and at the same time, the lower the overall cost of repairs. The simpler it is to disambiguate inadequate referring expressions in context, the lower the cost of any individual failure should be. The JCT should absorb misinterpretations cheaply: no one dies if an error is made, the interlocutors have a good acoustic channel and share a task goal, the universe of discourse is small (objects on the game screen rather than the unknown location of an emergency) with familiar, readily codable distinctions (pink v red rather than normal v agonal breathing). A JCT speaker balances the cost of inferences based on what she does not see against the low risk of a low cost repair.

Since we found no overall effects of gaze- or mouse-cross-projection that cannot be reduced to a local effect, we assume that inferences about who can see what are drawn when necessary in the current experiment, and not used as longer term settings. However sensitive to listener details gestures have the potential to be (Bangerter, 2004; Galati & Brennan, 2013), here Different Role dyads often failed to draw inferences in time to edit their mouse gestures appropriately. Though pointing gestures are thought to originate with the linguistic expressions they accompany (McNeill, 1985; McNeill & Duncan, 2000; Morrel-Samuels & Krauss, 1992) they may launch even earlier than the speech itself (Bergmann, Aksu, & Kopp, 2011; De Ruiter, 2000) deriving as they do here from the deictic referring expression as a whole. If so, demonstration gestures may be difficult to intercept in response to later arriving inferences. Same Role dyads, we argue, have more capacity, while generating referring expressions, for more or earlier inference and for inference-based correction of gestures.

The gaze patterns of Same Role and Different Role dyads tell us that more than deixis may be involved. Same Role dyads, for example, key simpler forms to situations where on average the addressee has already accessed the referent. Either by prediction or by monitoring, Same Role speakers are better able to key the forms they use to their addressee's situation.

None of the major concepts in our account is really new. It combines least collaborative effort (Clark, et al., 1983; Clark & Wilkes-Gibbs, 1986) with perspective adjustment (P. Brown & Dell, 1987; Keysar, Barr, Balin, & Paek, 1998; Pickering & Garrod, 2004) in key cases, but with all important factors — perception-production priming, expected common ground, costs of acting on expectations — treated as variables, just as constraint satisfaction models (Brown-Schmidt, 2011) would suggest. By stressing the state of the interlocutors' production models and the need for inference in making some adjustments, our account makes it possible to distinguish between anticipated states and cognitive actions based on them, as Barr's (2008) Anticipation-Integration model does. Nonetheless, the model provides some linguistic alignment without separate cognitive costs: One ready explanation for the behaviour of Same Role dyads is that they had aligned models of expectancy of mention (Arnold, 2008).

Most of the existing models, however, deal with the addressee's interpretation of referring expressions, while this paper examines the production of referring expressions and includes both speaker and addressee as part of the process. This work extends findings on interpretation by showing that common ground is not always respected by live interlocutors who need to communicate (Brown-Schmidt & Hanna, 2011). At the same time, the work challenges models of natural language generation to focus on accessibility of referents. Plainly there are first mentions which are not definite NPs and some may rely less on distinctive physical attributes of the referent than on it pragmatic qualities, —for example, whether it is physically in hand when first mentioned. There are already findings that suggest effects of codability of distinctions (Viethen, Goudbeek, & Krahmer, 2012) or set size (Gatt, van Gompel, Krahmer, & van Deemter, 2013), so that a metric for the difficulty of recovering from underspecifications should not be far away. Finally there is a model designed to balance precision in pointing gestures against linguistic detail in an accompanying referring expression (van der Sluis & Krahmer, 2007), by explicitly assigning costs to characteristics of the expression within a graph structure. Although there now appears to be no such inverse relationship to simulate (De Ruiter, Bangerter, & Dings, 2012), this model may serve as a first step in constructing a system that reduces the complexity of referring expressions from their theoretical maxima (indefinite NPs with attributes and nouns) in response to increasing support from apparent common ground, decreasing risk and cost of repair, and increasing cost of genuinely monitoring for common ground given the extent to which speaker and addressee can benefit from prediction-production symmetries.

REFERENCES

- (2006, May 8, 2006). Channeltrans, from ftp://ftp.icsi.berkeley.edu/pub/speech/download/channeltrans/

Anderson, A., Bard, E., Sotillo, C., Newlands, A., & Doherty-Sneddon, G. (1997). Limited visual control of the intelligibility of speech in face-to-face dialogue. *Perception and Psychophysics, 59*, 580-592.

Ariel, M. (1988). Referring and accessibility. *Journal of Linguistics, 24*, 65-87.

Ariel, M. (1990). *Accessing Noun-Phrase Antecedents*. London: Routledge.

Ariel, M. (2001). Accessibility theory: An overview. In T. Sanders, J. Schilperoord & W. Spooren (Eds.), *Text representation: Linguistic and psycholinguistic aspects* (pp. 29-87). Amsterdam: John Benjamins.

Arnold, J. (2001). The effect of thematic roles on pronoun use and frequency of reference continuation. . *Discourse Processes, 31*, 137–162.

Arnold, J. (2008). Reference production: Production-internal and addressee-oriented processe. *Language and Cognitive Processes, 23*, 495-527.

Arnold, J., & Griffin, Z. (2007). The effect of additional characters on choice of referring expression: Everyone competes. *Journal of Memory and Language, 56*, 521-536.

Bangerter, A. (2004). Using pointing and describing to achieve joint focus of attention in dialogue. *Psychological Science, 15*(6), 415-419.

Bard, E., Anderson, A., Chen, Y., Nicholson, H., Havard, C., & Dalzel-Job, S. (2007). Let's you do that: Sharing the cognitive burdens of dialogue. *Journal of Memory and Language, 57*(4), 616-641.

Bard, E., Anderson, A., Sotillo, C., Aylett, M., Doherty-Sneddon, G., & Newlands, A. (2000). Controlling the intelligibility of referring expressions in dialogue *Journal of Memory and Language, 42* 1-22.

Bard, E., & Aylett, M. (2004). Referential form, word duration, and modeling the listener in spoken dialogue. In J. C. Trueswell & M. K. Tanenhaus (Eds.), *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions.* (pp. 173-191). Cambridge, MA: MIT Press.

Barr, D. (2008). Pragmatic expectations and linguistic evidence: Listeners anticipate but do not integrate common ground. *Cognition, 109*, 18-40.

Barr, D., & Keysar, B. (2002). Anchoring comprehension in linguistic precedents. *Journal of Memory and Language, 46*, 391-418.

Bergmann, K., Aksu, V., & Kopp, S. (2011). *The relation of speech and gestures: Temporal synchrony follows semantic synchrony.* Paper presented at the Proceedings of the Second Workshop on Gesture and Speech in Interaction (GeSpin 2011).

Brennan, S. (2005). How conversation is shaped by visual and spoken evidence. In J. Trueswell & M. Tanenhaus (Eds.), *Approaches to studying world-situated language use: Bridging the language-as-product and language-action traditions* (pp. 95-129). Cambridge, MA: MIT Press.

Brennan, S., Chen, X., Dickinson, C., Neider, M., & Zelinsky, G. (2008). Coordinating cognition: The costs and benefits of shared gaze during collaborative search. *Cognition, 106*(3), 1465-1477.

Brennan, S., & Clark, H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory and Cognition, 11*, 1482-1493.

Brown, P., & Dell, G. (1987). Adapting production to comprehension: The explicit mention of instruments. *Cognitive Psychology, 19*.

Brown, R. (1958). How shall a thing be called? *Psychological Review, 65*, 14-21.

Brown-Schmidt, S. (2009). Partner-specific interpretation of maintained referential precedents during interactive dialog. *Journal of Memory and Language, 61*, 171–190.

Brown-Schmidt, S. (2011). Beyond common and privileged: Gradient representations of common ground in real-time language use. *Language and Cognitive Processes, 27*, 62-89.

Brown-Schmidt, S., Campana, E., & Tanenhaus, M. (2004). Real-time reference resolution by naive participants during a task-based unscripted conversation. In J. Trueswell & M. Tanenhaus (Eds.), *Approaches to Studying World-Situated Language Use* (pp. 153-171). Cambridge, MA: MIT Press.

Brown-Schmidt, S., & Hanna, J. (2011). Talking in another person's shoes: Incremental perspective-taking in language processing. *Dialogue and Discourse 2* 11-33.

Carletta, J., Hill, R., Nicol, C., Taylor, T., De Ruiter, J.-P., & Bard, E. (2010). Eyetracking for two-person tasks with manipulation of a virtual world. *Behavior Research Methods, 42*(1), 254-265.

Carletta, J., & Mellish, C. (1996). Risk-taking and recovery in task-oriented dialogue. *Journal of Pragmatics, 26*, 71–107.

Chartrand, T., & Bargh, J. (1999). The Chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology, 76*, 893-910.

Clark, H., & Marshall, C. (1981). Definite reference and mutual knowledge. In A. Joshi, B. Webber & I. Sag (Eds.), *Elements of discourse understanding* (pp. 10-63). Cambridge: Cambridge University Press.

Clark, H., Schreuder, R., & Buttrick, S. (1983). Common ground and the understanding of demonstrative reference. *Journal of Verbal Learning and Verbal Behavior*(22), 245-258.

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

Dale, R., & Reiter, E. (1995). Computational interpretations of the Gricean Maxims in the generation of referring expressions. *Cognitive Science, 18*, 233-263.

Dale, R., & Viethen, J. (2010). Attribute-centric referring expression generation. In E. Krahmer & M. Theune (Eds.), *Empirical Methods in Natural Language Generation* (Vol. Springer, pp. 163–179).

De Ruiter, J.-P. (2000). The production of gesture and speech. In D. McNeill (Ed.), *Language and Gesture* (pp. 284-311). Cambridge, UK: Cambridge University Press.

De Ruiter, J.-P., Bangerter, A., & Dings, P. (2012). The interplay between gesture and speech in the production of referring expressions: Investigating the Tradeoff Hypothesis. *Topics in Cognitive Science 4* 232–248.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition, 109*(2), 193-210.

Doherty-Sneddon, G., Bruce, V., Bonner, L., Longbotham, S., & Doyle, C. (2002). Development of gaze aversion as disengagement from visual information. *Developmental Psychology, 38*, 483-485.

Ferreira, V., & Hudson, M. (2011). Saying “that” in dialogue: The influence of accessibility and social factors on syntactic production. *Language and Cognitive Processes, 26*(10), 1736-1762.

Fillmore, C. (1982). Towards a descriptive framework for spatial deixis. In R. Jarvella & W. Klein (Eds.), *Speech, place, and action*. New York: Wiley.

Foster, M.-E., Bard, E., Guhe, M., Hill, R., Oberlander, J., & Knoll, A. (2008). *The roles of haptic-ostensive referring expressions in cooperative task-based human-robot dialogue*. Paper presented at the Third ACM/IEEE International Conference on Human Robot Interaction (HRI 2008).

Fukumura, K., & van Gompel, R. (2010). Choosing anaphoric expressions: Do people take into account likelihood of reference? *Journal of Memory and Language, 62*, 52-66.

Fukumura, K., & Van Gompel, R. (2012). Producing pronouns and definite noun phrases: Do speakers use the addressee's discourse model? *Cognitive Science, 36*, 1289-1311.

Fukumura, K., van Gompel, R., & Pickering, M. (2010). The use of visual context during the production of referring expressions. *The Quarterly Journal of Experimental Psychology, 63*, 1700–1715.

Galati, A., & Brennan, S. (2010). Attenuation information in spoken communication: For the speaker, or for the addressee? *Journal of Memory and Language, 62*, 35-51.

Galati, A., & Brennan, S. (2013). Speakers adapt gestures to addressees' knowledge: Implications for models of co-speech gesture. *Language and Cognitive Processes*.

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends In Cognitive Sciences, 8*(1), 8-11.

Gatt, A., van Gompel, R., Krahmer, E., & van Deemter, K. (2013). *Models and empirical data for the production of referring expressions: The case of domain size and content selection.*Unpublished manuscript.

Grosz, B., Joshi, A., & Weinstein, S. (1995). Centering: A Framework for modeling the local coherence of discourse *Computational Linguistics, 21*, 203-225.

Gundel, J., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language, 69*, 274-307.

Gundel, J., Hedberg, N., & Zacharski, R. (2012). Underspecification of cognitive status in reference production: Some empirical predictions. *Topics in Cognitive Science*.

Hanna, J., & Brennan, S. (2007). Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language, 57*, 596-615.

Hanna, J., & Tanenhaus, M. (2004). Pragmatic effects on reference resolution in a collaborative task: evidence from eye movements. *Cognitive Science, 28*(1), 105-115.

Hanna, J., Tanenhaus, M., & Trueswell, J. (2003). The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language, 49*(1), 43-61.

Horton, W., & Gerrig, R. (2002). Speakers' experiences and audience design: knowing when and knowing how to adjust utterances to addressees. *Journal Of Memory And Language, 47*(4), 589-606.

Horton, W., & Gerrig, R. (2005a). Conversational common ground and memory processes in language production. *Discourse Processes, 40*(1), 1-35.

Horton, W., & Gerrig, R. (2005b). The impact of memory demands on audience design during language production. *Cognition, 96*(2), 127-142.

Horton, W., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition, 59*, 91-117.

Jesse, A., & Johnson, E. (2012). Prosodic temporal alignment of co-speech gestures to speech facilitates referent resolution. *Journal Of Experimental Psychology-Human Perception And Performance, 38*(6), 1567-1581.

Keysar, B., Barr, D., Balin, J., & Paek, T. (1998). Definite reference and mutual knowledge: Process models of common ground in comprehension. *Journal of Memory and Language, 39*(1), 1-20.

Krahmer, E., & van Deemter, K. (2012). Computational generation of referring expressions: A survey. *Computational Linguistics, 38*, 173-218.

Kronmüller, E., & Barr, D. (2007). Perspective-free pragmatics: Broken precedents and the recovery-from-preemption hypothesis. *Journal Of Memory And Language*(56), 436-455.

Levy, R. (2008). Expectation-based syntactic comprehension. . *Cognition, 106*, 1126–1177.

Levy, R., Fedorenko, E., Breen, M., & Gibson, E. (2012). The processing of extraposed structures in English. *Cognition, 122*, 12-36.

Lindblom, B. (1990). Explaining variation: a sketch of the H and H theory. In W. Hardcastle & A. Marchal (Eds.), *Speech Production and Speech Modelling*. Dordrecht, Netherlands: Kluwer.

Lobmaier, J., Fischer, M., & Schwaninger, A. (2006). Objects capture perceived gaze direction. *Experimental Psychology, 53*(2), 117-122.

Louwerse, M., Dale, R., Bard, E., & Jeuniaux, P. (2012). Behavior matching in multimodal communication is synchronized *Cognitive Science, 36*(8), 1404-1426.

Lyons, J. (1977). *Semantics* (Vol. 2). Cambridge: Cambridge University Press.

McNeill, D. (1985). So you think gestures are nonverbal? . *Psychological Review, 92*, 350–371.

McNeill, D., & Duncan, S. (2000). Growth points in thinking-for-speaking. In D. McNeill (Ed.), *Language and gesture* (pp. 141–161). Cambridge, UK: Cambridge University Press.

Meltzoff, A., & Moore, M. (1977). Imitation of facial and manual gestures by human neonates *Science, 198*, 75-78.

Metzing, C., & Brennan, S. (2003). When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal Of Memory And Language, 49*(2), 201-213.

Meyer, A., van der Meulen, F., & Brooks, A. (2004). Eye movements during speech planning: Talking about present and remembered objects. *Visual Cognition, 11*(5), 553-576.

Morrel-Samuels, P., & Krauss, R. (1992). Word familiarity predicts the temporal asynchrony of hand gestures and speech. *Journal of Experimental Psychology: Learning, Memory and Cognition, 18*, 615-623.

Pickering, M., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences, 27*(2), 169-190.

Pickering, M., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends In Cognitive Sciences, 11*(3), 105-110.

Pietsch, C., Buch, A., Kopp, S., & de Ruiter, J.-P. (2012). Measuring syntactic priming in dialogue corpora. In B. Stolterfoht & S. Featherston (Eds.), *Empirical Approaches to Linguistic Theory: Studies in Meaning and Structure* (pp. 29 - 42.). Berlin Mouton de Gruyter.

Potter, M., & Lombardi, L. (1998). Syntactic priming in immediate recall of sentences. *Journal of Memory and Language, 38*, 265–282.

Prince, E. (1981). Toward a taxonomy of given-new information. In P. Cole (Ed.), *Radical Pragmatics* (pp. 223-256). New York: Academic press.

Richardson, D., & Dale, R. (2005). Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science, 29*(6), 1045-1060.

Richardson, D., Dale, R., & Kirkham, N. (2007). The art of conversation is coordination: common ground and the coupling of eye movements during dialogue. *Psychological Science, 18*(5), 407-413.

Richardson, M., Marsh, K., & Schmidt, R. (2005). Effects of visual and verbal interaction on unintentional interpersonal coordination. *Journal of Experimental Psychology: Human Perception and Performance, 31*(1), 62-79.

Rosa, E., & Arnold, J. (2011). *The role of attention in choice of referring expressions.* Paper presented at the PRE-CogSci: Briging the gap between computational, empirical and theoretical approaches to reference, Boston.

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*, 269-273.

Shockley, K., Santana, M. V., & Fowler, C. A. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology: Human Perception and Performance, 29*(2), 326-332.

Smith, M., & Wheeldon, L. (2001). Syntactic priming in spoken sentence production - an online study. *Cognition, 78*(2), 123-164.

Smith, S., Noda, H., Andrews, S., & Jucker, A. (2005). Setting the stage: How speakers prepare listeners for the introduction of referents in dialogues and monologues. *Journal of Pragmatics, 37*, 1865-1895.

van Deemter, K., Gatt, A., van Gompel, R., & Krahmer, E. (2012). Towards a computational psycholinguistics of reference production. *Topics in Cognitive Science, 4*(2), 166-183.

van der Sluis, I., & Krahmer, E. (2007). Generating multimodal referring expressions. *Discourse Processes, 44*(3), 145-174.

Viethen, J., Goudbeek, M., & Krahmer, E. (2012, August, 2012). *The impact of colour difference and colour codability of reference production.* Paper presented at the The 34th annual meeting of the Cognitive Science Society (CogSci), Sapporo, Japan.

Vogels, J., Krahmer, E., & Maes, A. (2012). Who is where referred to how, and why? The influence of visual saliency on referent accessibility in spoken language production. *Language and Cognitive Processes*.

Zbilut, J. P., Giuliani, A., & Webber, C. L. (1998). Detecting deterministic signals in exceptionally noisy environments using cross-recurrence quantification. *Physics Letters A, 246*(1-2), 122-128.

TABLES

Table 1.

Accessibility Coding Scheme

Coding Level	Form of Referring Expression	Examples
0 Indefinites	Indefinite NP	*a purple one* *one of the nearest blue pieces*
	Bare nominal	*pink one* *triangles*
1 Definites	Definite NP	*the red bit* *the other purple one*
2 Deictics	Deictic NP	*those two little kids.*
	Deictic Possessive } Pronouns	*these* *mine*
3 Pronouns	Other Pronouns	*it*
	Clitic/inaudible.	*-/z/*

Table 2.

Global effects of Gaze Cross-Projection, Mouse Cross-Projection, and Roles Assigned on Distribution of Initial Mentions across Forms of Referring Expression

	Forms of Referring Expression					
Predictors	Definites		Deictics		Pronouns	
Predictors	B	Wald	B	Wald	B	Wald
Intercept	.386	7.105	.802	33.032	-.409	5.129
Roles assigned	.455	9.644‡	.282	3.669#	.264	1.998
Gaze cross-projection	-.068	.219	.025	.031	-.064	.120
Mouse cross-projection	.577	15.761§	-.249	2.931#	.076	.1710

Note. Multinomial Logistic Regression with backwards elimination of Player and interactions.

# *p* < .10, * *p* < .05; ‡ *p* < .01; § *p* < .001

Table 3.

Local Effects of Speaker's Actions and Roles Assigned on the Distribution of Initial Mentions Across Forms of Referring Expression 3a). Show Mouse, 3b) No Mouse

3a. Show Mouse

	Forms of Referring Expression					
	Definites		Deictics		Pronouns	
Predictors	B	Wald	B	Wald	B	Wald
Intercept	.425	2.728	1.359	35.726	-1.111	7.816‡
Speaker move	-.252	.938	-.713	8.665‡	-.332	1.037
Roles assigned	.587	2.384	.195	.304	.574	.992
Speaker hover	.223	.505	-.247	.725	1.219	7.229‡
Speaker hover * Roles assigned	-.345	.553	.352	.649	-.435	.442

Note. Multinomial Logistic Regression with backwards elimination of interactions other than Roles Assigned x Speaker Hover for Mouse Cross-Projection conditions Show Mouse and No Mouse.

* $p < .05$; ‡ $p < .01$; § $p < .001$

3b. No Mouse

Predictors	Forms of Referring Expression					
	Definites		Deictics		Pronouns	
	B	Wald	B	Wald	B	Wald
Intercept	1.838	38.308	1.869	38.646	-.626	1.839
Speaker move	-.186	.565	-.770	8.888‡	-.140	.174
Roles assigned	-.314	.773	-.648	3.042	-.248	.180
Speaker hover	-1.191	13.191§	-1.117	10.727‡	.439	.797
Speaker hover * Roles assigned	1.210	7.687‡	1.200	6.740‡	.794	1.457

Note. Multinomial Logistic Regression with backwards elimination of interactions other than Roles Assigned x Speaker Hover for Mouse Cross-Projection conditions Show Mouse and No Mouse.

* $p < .05$; ‡ $p < .01$; § $p < .001$

Table 4.

Local Effects of Speaker's Actions on the Distribution of Initial Mentions Across Forms of Referring Expression for Different Role Dyads with No Mouse Cross-Projection.

	Forms of Referring Expression					
	Definites		Deictics		Pronouns	
Predictors	B	Wald	B	Wald	B	Wald
Intercept	1.833	21.054	1.670	16.742	-.580	.842
Player	.314	.316	.573	1.017	.475	.293
Speaker move	-.436	1.455	-.844	5.204*	1.048	2.419
Speaker hover	-.823	3.647#	-.773	2.888#	-.848	.782
Speaker hover * Player	-.614	.912	-.705	1.123	-.848	.293

Note. Multinomial Logistic Regression with backwards elimination of interactions other than Player x Speaker Hover.

# *p* < .10 ; * *p* < .05; ‡ *p* < .01; § *p* < .001

Table 5.

Gaze Cross-Recurrence: Distribution of Analyzed Referring Expressions by Accessibility and Roles of Speakers

	Accessibility Category of Referring Expression			
	0	1	2	3
Roles Assigned	Indefinites	Definites	Deictics	Pronouns
Same	46	168	152	57
Different	89	162	187	75

FIGURES

Figure 1. A Joint Construction Task screen showing static features (target, new parts set, breakage count, timer) and movable parts (original parts, viewer's and collaborator's mice, and collaborator's gaze).

Figure 2. Demonstration with deictics as seen by the addressee in the Joint Construction Task. The smaller square is a mouse cursor with a fill colour indicating whether the speaker has selected the larger object. Figures 2a and 2b (moving the referent part) and 2c (merely superimposing, a mouse cursor over it (‘hovering’) when the cursor is cross-projected to the addressee’s screen) show actions visible to the addressee. In Figure 2d (hovering when the speaker’s mouse cursor is not cross-projected) is an action privileged to the speaker, which cannot help the listener interpret ‘that square’.

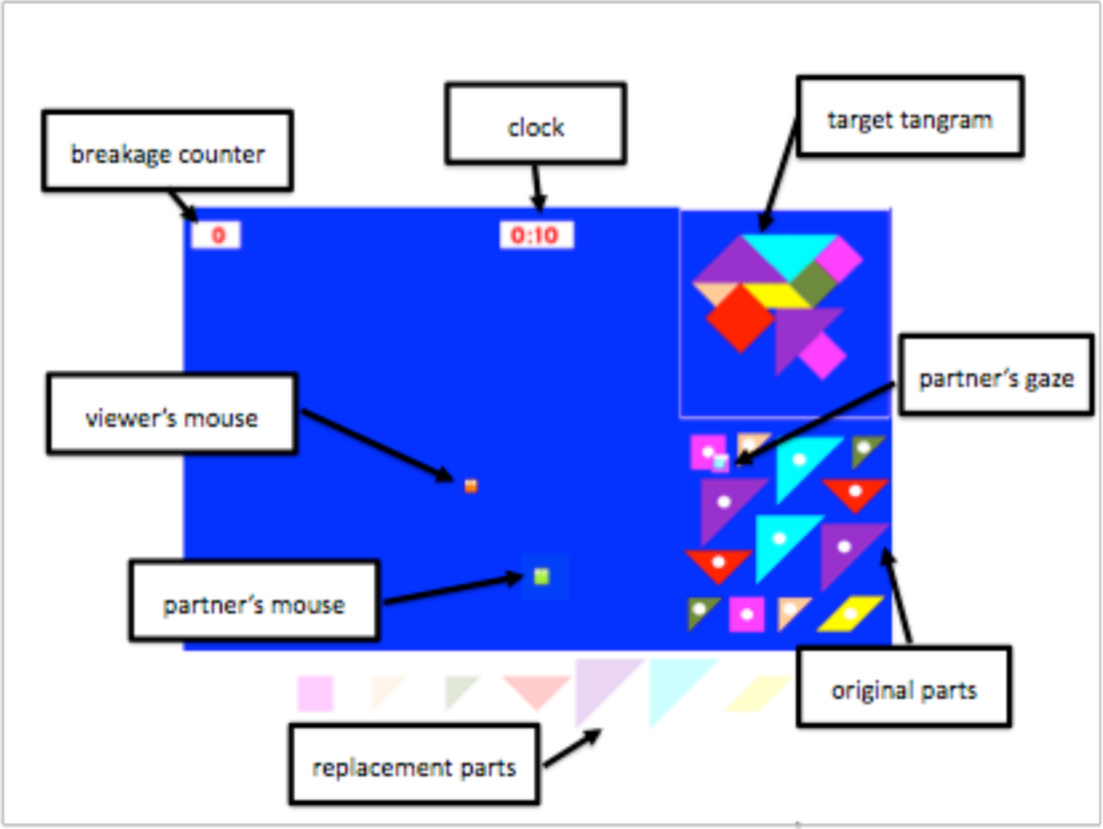
Figure 3. Changes in the distribution of first mentions when the speaker moves the referent. (3a) Show Mouse trials, mouse cursors cross-projected, (3b) No Mouse trials, mouse cursors not cross-projected. In each case the left panel shows probabilties and the right panel shows odds ratios relative to the base category, indefinites.

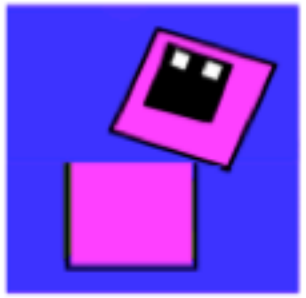
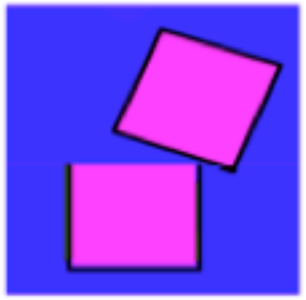
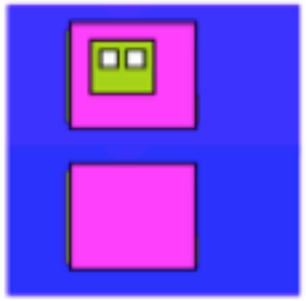
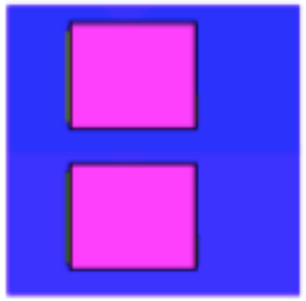
Figure 4. Changes in the distribution of first mentions when the speaker hovers the mouse over the referent without moving it, by Roles Assigned (4a) Show Mouse trials, mouse cursors cross-projected, (4b) No Mouse trials, mouse cursors not cross-projected. In each case the left panel shows probabilties and the right panel shows odds ratios relative to the base category, indefinites.

Figure 5 Examples of synchrony and near synchrony in a dyad’s actions. N and G are imaginary individuals who synchronize their actions well. V and S synchronize imperfectly.

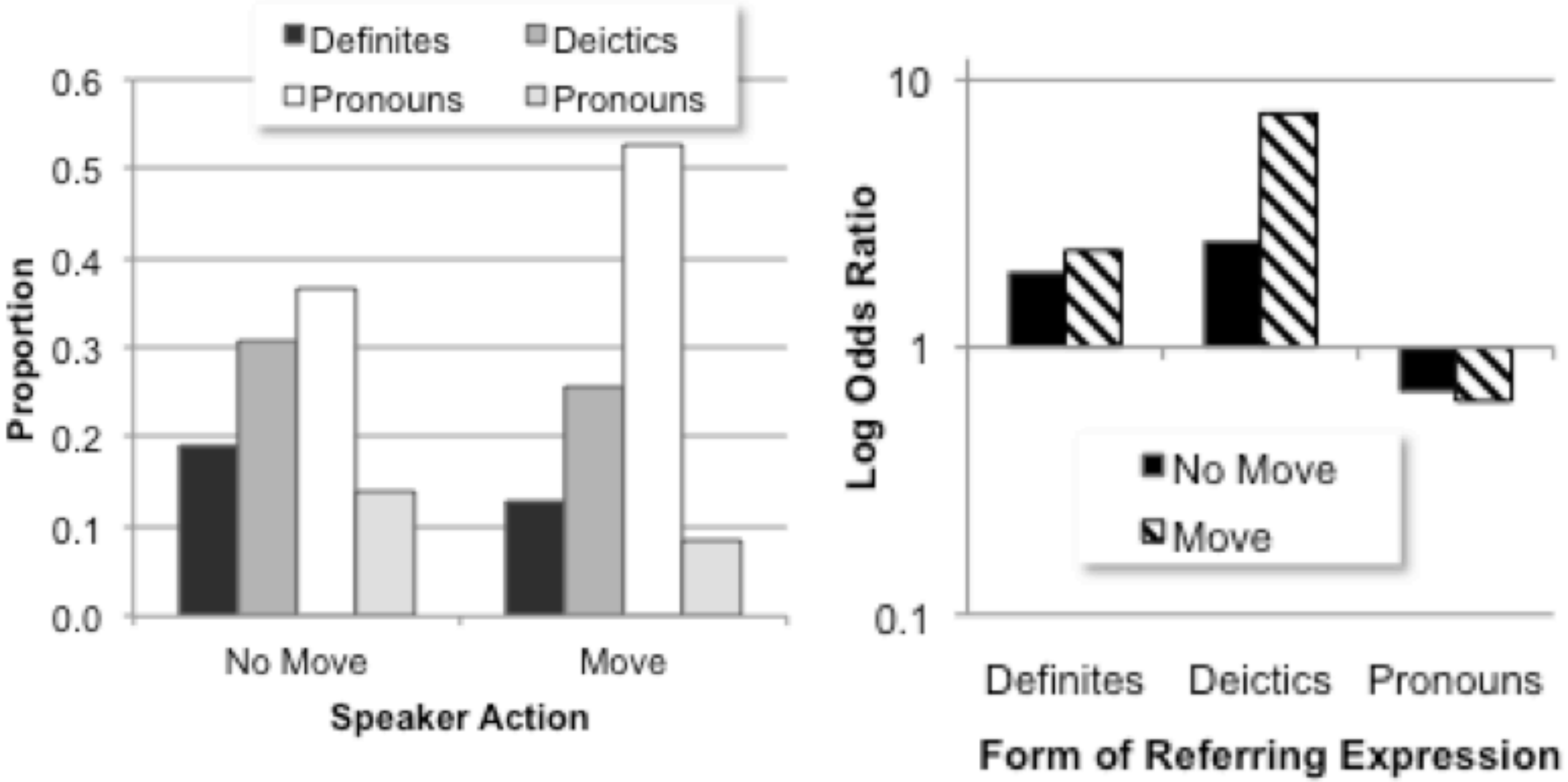
Figure 6. Real and random cross-recurrence of players' gaze around the onset of initial mentions of each referential form; a) for Same Roles dyads; b) for Different Roles dyads.

Figure 7. Order of interlocutors' gaze with different forms of referring expression: Mean recurrent gaze probabilities on the negative lags (listener first) less those on the positive lags (speaker first) by dyad role assignments.

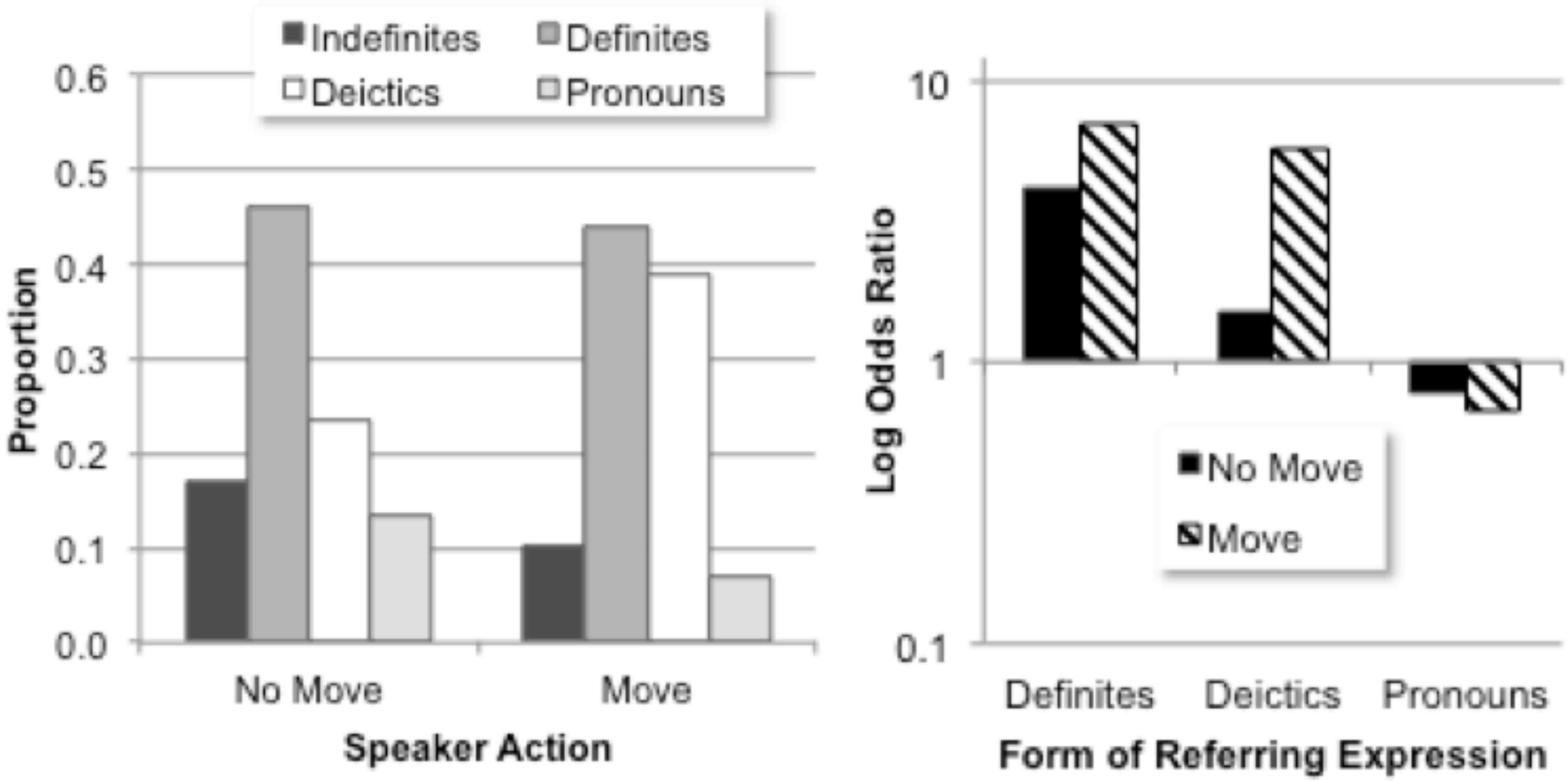

breakage counter
clock
target tangram
0
0:10
viewer's mouse
partner's gaze
partner's mouse
original parts
replacement parts

Concurrent Action	Cross-Projection of Mouse Cursor		Referring Expression
	Show Mouse	No Mouse	
Moving	2a) 	2b) 	*‘This square’*
Hovering	2c) 	2d) 	

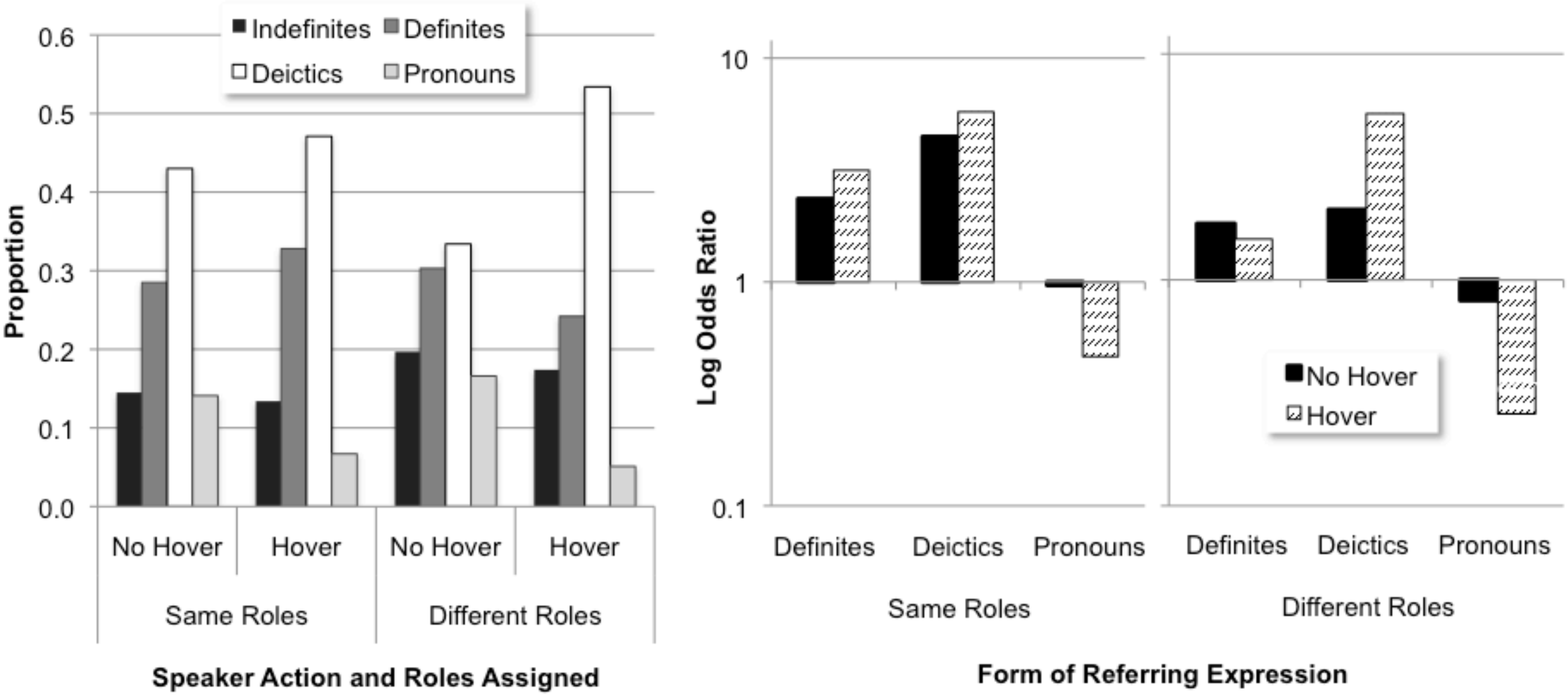
3a)


■ Definites
■ Deictics
□ Pronouns
■ Pronouns
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Proportion
No Move
Move
Speaker Action
10
1
0.1
Log Odds Ratio
■ No Move
▨ Move
Definites
Deictics
Pronouns
Form of Referring Expression

3b)

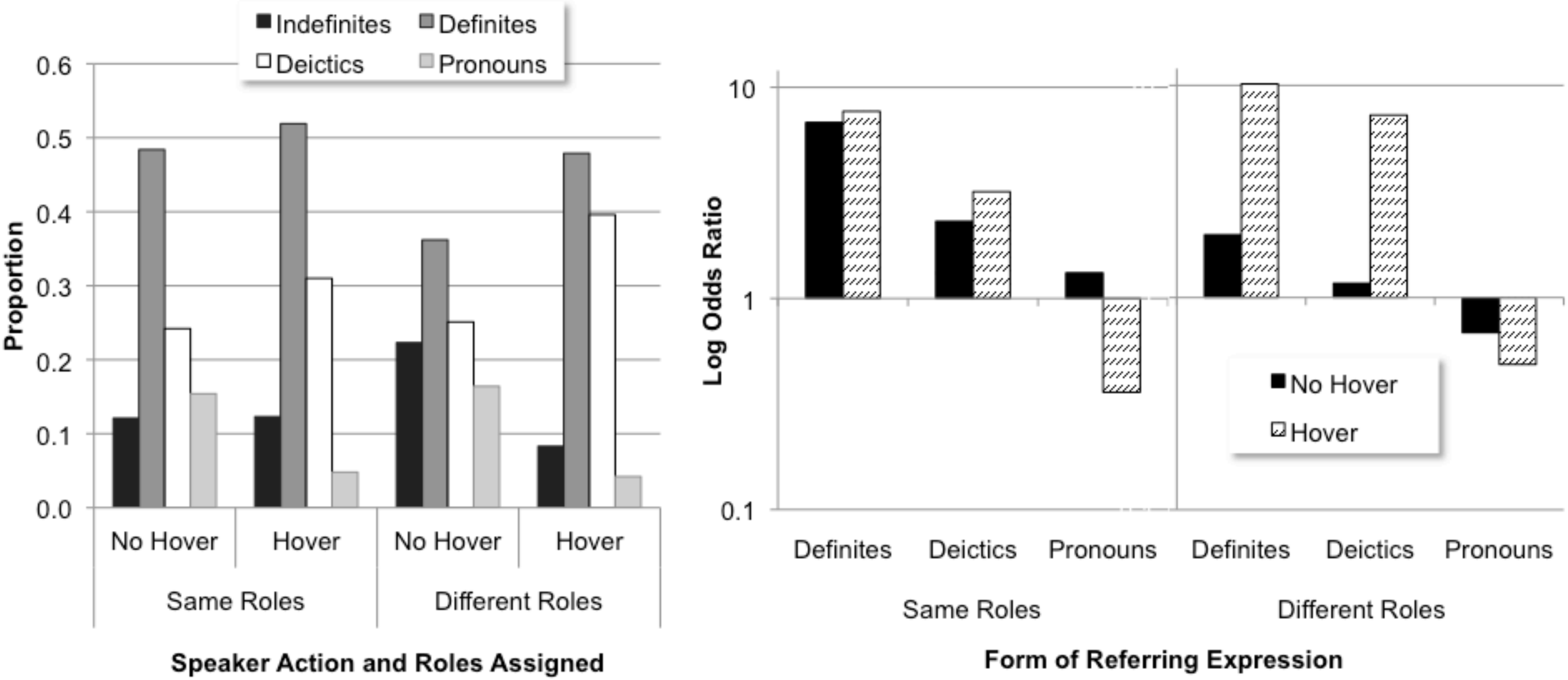

Indefinites
Definites
Deictics
Pronouns
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Proportion
No Move
Move
Speaker Action
10
1
0.1
Log Odds Ratio
No Move
Move
Definites
Deictics
Pronouns
Form of Referring Expression

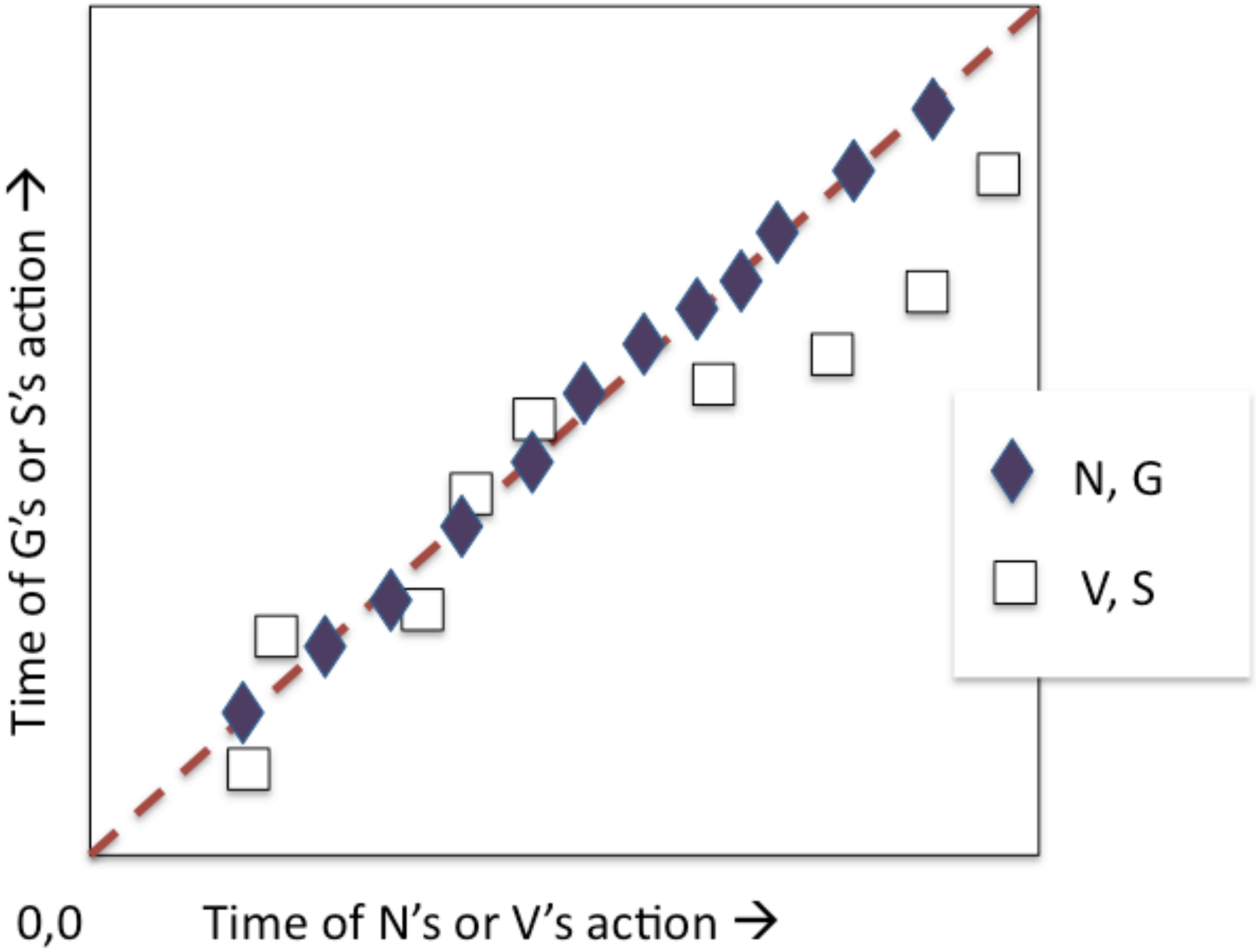
4a

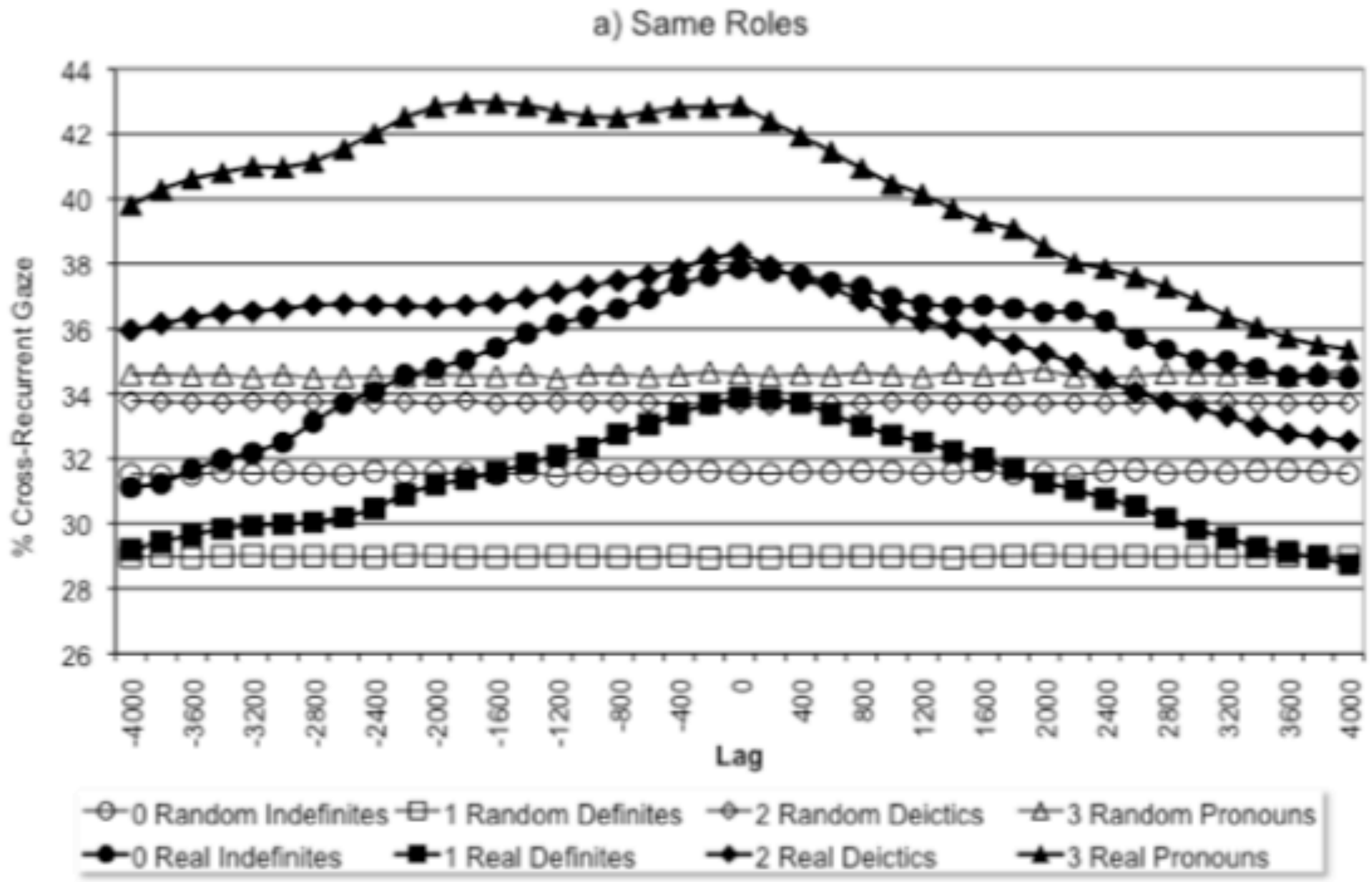


Indefinites
Definites
Deictics
Pronouns
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Proportion
No Hover
Hover
No Hover
Hover
Same Roles
Different Roles
Speaker Action and Roles Assigned
10
1
0.1
Log Odds Ratio
No Hover
Hover
Definites
Deictics
Pronouns
Definites
Deictics
Pronouns
Same Roles
Different Roles
Form of Referring Expression

4b

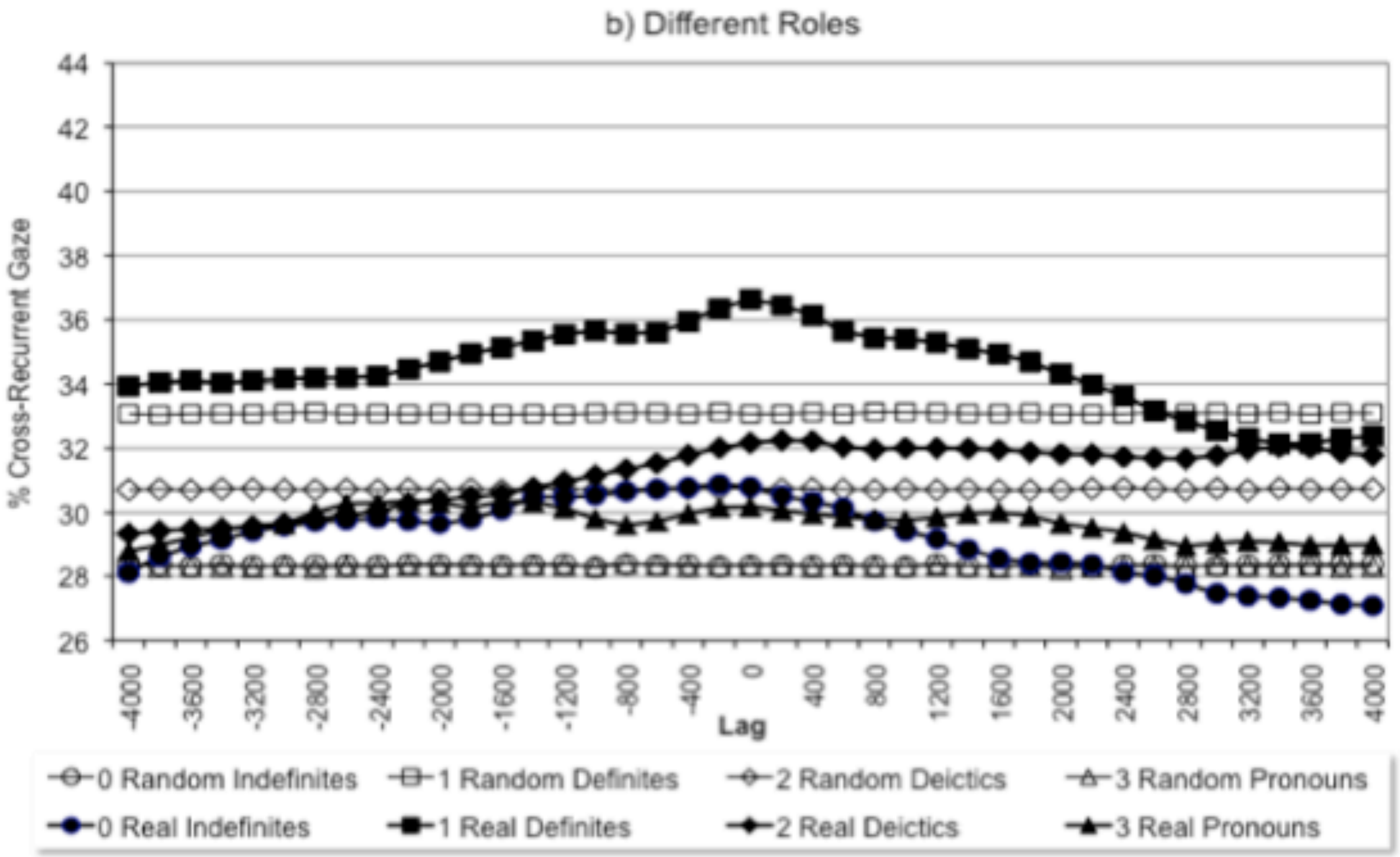

■ Indefinites
■ Definites
□ Deictics
■ Pronouns
Proportion
0.6
0.5
0.4
0.3
0.2
0.1
0.0
No Hover
Hover
No Hover
Hover
Same Roles
Different Roles
Speaker Action and Roles Assigned
Log Odds Ratio
10
1
0.1
■ No Hover
▨ Hover
Definites
Deictics
Pronouns
Definites
Deictics
Pronouns
Same Roles
Different Roles
Form of Referring Expression

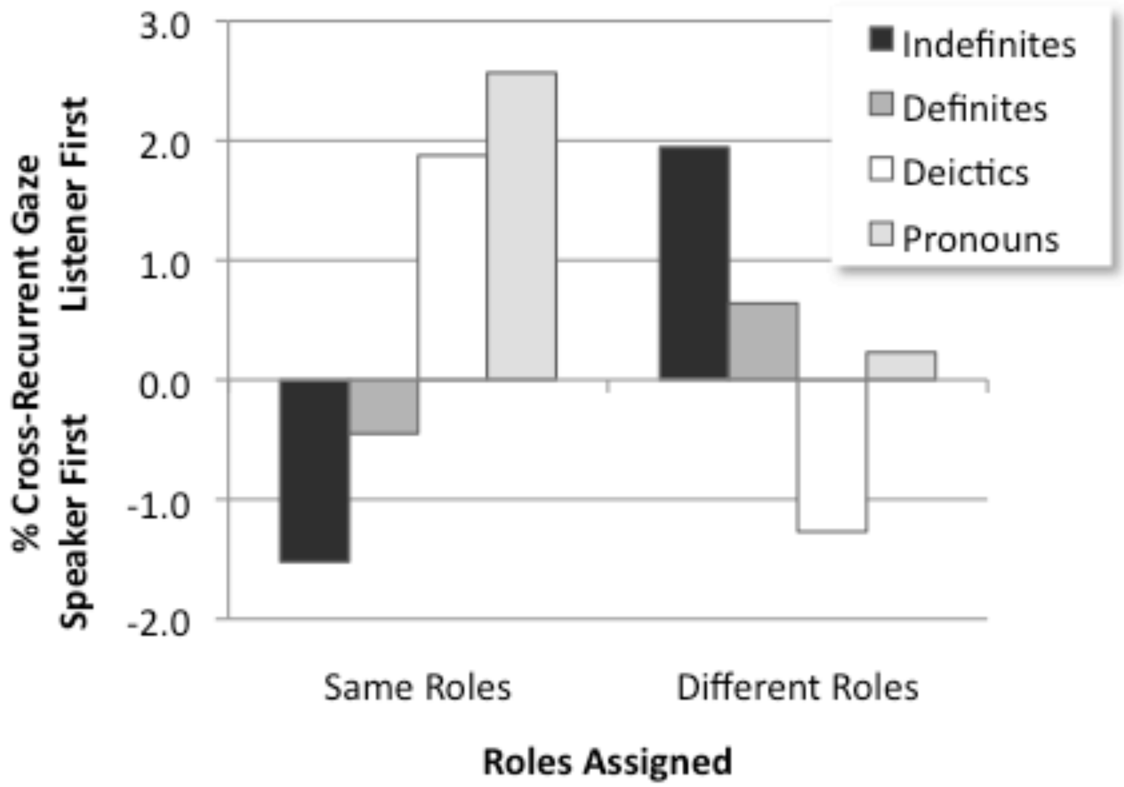

Time of G's or S's action →
N, G
V, S
0,0
Time of N's or V's action →



a) Same Roles

% Cross-Recurrent Gaze
44
42
40
38
36
34
32
30
28
26
-4000
-3600
-3200
-2800
-2400
-2000
-1600
-1200
-800
-400
0
400
800
1200
1600
2000
2400
2800
3200
3600
4000
Lag
0 Random Indefinites
1 Random Definites
2 Random Deictics
3 Random Pronouns
0 Real Indefinites
1 Real Definites
2 Real Deictics
3 Real Pronouns


b) Different Roles



% Cross-Recurrent Gaze
Listener First
Speaker First
3.0
2.0
1.0
0.0
-1.0
-2.0
Indefinites
Definites
Deictics
Pronouns
Same Roles
Different Roles
Roles Assigned