Accessibility and Attention in Situated Dialogue: Roles and Regulations

Ellen Gurman Bard (ellen@ling.ed.ac.uk)
Linguistics and English Language, University of Edinburgh, Edinburgh, UK

Robin Hill (r.l.hill@ed.ac.uk), Manabu Arai (manabu.arai@ed.ac.uk),
Human Communication Research Centre, University of Edinburgh, Edinburgh, UK

Mary Ellen Foster (M.E.Foster@ed.ac.uk)
Informatik VI: Robots and Embedded Systems, Technische Universität, München, Germany

Abstract

Accessibility theory (Ariel, 1988; Gundel, Hedberg, & Zacharski, 1993) proposes that the grammatical form of referring expression depends on the accessibility of its referent, with greater accessibility permitting more reduced expressions. This paper presents evidence discriminating between Realist and Declarative views of how accessibility will be determined. First, it presents a corpus study of first mentions of on-screen objects produced during a joint construction task. As the Declarative view suggests, elaboration of referring expressions is not controlled only by real local circumstances but instead differs with the assigned ability of interlocutors to declare how the dialogue will proceed. Second, it shows that speakers' visual attention around such expressions supports the same model. The context of a joint physical task does automatically align players' attention. Instead, players who declare accessibility inappropriate to their listeners' perspectives align attention poorly, while those whose expressions show better design, have gaze patterns predictable from accessibility theory.

Keywords: reference, accessibility, corpus experimental studies, pragmatics, situated dialogue, cross-recurrent gaze

Introduction

The question "How shall a thing be called?" (Brown, 1958) still engages anyone who deals with language production. One very wide-ranging approach (Ariel, 1988, 1990, 2001) relates elaboration of the form of referring expressions to how difficult the producer estimates it will be to access the referent concept or entity. Expressions introducing entities deemed completely unfamiliar to the audience could be as complex as indefinite NPs with modifiers of various kinds. Expressions for referents or antecedents of intermediate accessibility might be simple definite NPs, deictics, or personal pronouns in that order. Expressions for the most recent focussed antecedent can be as minimal as cliticized pronouns *(/z/ in the garage*). Though Ariel (2001) concentrates on the producer's estimate for the accessibility of a text antecedent, the system is general enough to apply in the interactive settings used to study natural language generation and human dialogue.

Accessibility theory offers a unified framework for predicting how forms of referring expression will respond to many conditions which might draw attention to the correct referent. Our research asks whose attempt to find a referent determines referential form, and whether, if speaker and listener are co-present, there is any point in attempting to distinguish between them.

We consider two main approaches to referring expression design. Most current approaches could be described as Realist: They treat the design of referring expressions as a reflex of genuine facts local to the expression in time or space. These approaches differ in the set of local facts which should bear on the speaker's choice. To be maximally felicitous, referring expressions should reflect the accessibility of the intended referent from the audience's perspective. Alternatively, referring expressions may reflect accessibility from the speaker's point of view. A burgeoning literature tends to support Realist audience-oriented design for features which endure throughout a dialogue (Bard et al., 2007; Brennan, Chen, Dickinson, Neider, & Zelinsky, 2007; Brennan & Clark, 1996), but reveals very mixed sensitivity to more incremental changes in audience knowledge (Bard et al., 2007; Bard et al., 2000; Bard & Aylett, 2004; Brown-Schmidt, Campana, & Tanenhaus, 2004; Horton & Gerrig, 2002, 2005a, 2005b; Horton & Keysar, 1996).

In contrast, what we call the Declarative approach proposes that the form of an expression is not controlled by local facts, but instead marks the producer's decision to treat the referent as accessible to a particular degree. The approach is inspired by Steedman's (2007) treatment of phenomena like accenting, de-accenting and contrastive stress, which can offer phonological analogues to elaboration of form. To our knowledge, this view has not been developed before for accessibility. Yet, just as a speaker has an option to express a contrast prosodically (*from the red triangle to the green one*) or not, s/he has an option to treat an entity as a clearly in focus, highly accessible *it* or not. Gundel, Hedberg and Zacharski (1993) account for part of this latitude by proposing that more accessible referents meet the criteria for felicitous use of all lower accessibility forms (*it / the triangle*). We suggest that there is also latitude to use higher accessibility forms when only the lower forms are justified by the facts available to audience. The option allows authors, for example, to convey to unknown readers who have no such prior understanding that there is a single entity predictable from its context (*the noticeboard and row of coathooks)*. For a discussion of how dialogues deal with such instances, see Gundel, Hedberg, and Zacharski (2001)

and Clark and Wilkes-Gibbs (1986).

The present work takes two steps towards establishing how accessibility is determined. First, it examines the accessibility of expressions produced under manipulations of speaker and listener knowledge and role in unscripted, situated collaborative dialogues. Second, it confirms that this setting is useful for testing accessibility.

Referring Expressions in Situ

To test Realist and Declarative predictions, we use a corpus of unscripted dialogues produced during a joint construction task carried out on yoked computer screens. The dependent variable is the distribution across levels of accessibility of the first mentions of movable on-screen objects.

We test the competing Realist positions by manipulating how much each speaker knows about the other's attention and actions, and by selecting expressions concurrent with actions which could enhance accessibility of referents. If the addressee's attention is important to the choice of referring expression, speakers able to track it should be able to use more accessible simple forms. For example, anything the addressee is currently holding or indicating could be 'that' rather than 'the red square'. Anything a speaker is indicating by movement or gesture could also attract deictic expressions (*this one*) (Foster et al., 2008; Kranstedt, Lücking, Pfeiffer, Rieser, & Wachsmuth, 2006): the change is audience-oriented if the listener can see such movements but speaker-oriented if they are private.

We test for Declarative control of referring expressions by manipulating the power to determine how a dialogue progresses: some pairs of players merely worked collaboratively; in others manager and assistant roles were assigned. Since managers can declare strategy and tactics, they should be prone to declaring accessibility. Since imbalance in roles makes the subsidiary player imitate the dominant one (Louwerse et al., 2009), assistants could act similarly. Declarative control of accessibility would produce more apparently speaker-oriented behavior in Manager-Assistant dialogues than in No Role dialogues. Realist control offers no reason to suppose that role assignment is important.

Method

Task. The Joint Construction Task (Carletta et al., submitted) offers two collaborating players an 11-part innominate target tangram, individual parts, spare parts, a work area, a breakage counter, and a timer (Figure 1). Players replicate the target tangram while maximizing speed and accuracy and minimizing cost in broken parts. Feedback on accuracy is provided at the end of each trial. As Figure 1 shows, the individual parts include 6 duplicated shape-color combinations and one singleton.

Players manipulate and rotate objects by mouse button and movement. Color distinguishes the viewer's and collaborator's mouse cursors, and the viewer's mouse when grasping an object. A circle shows were the collaborator is looking.

Objects can be joined only if each is held by a different player. Any object grasped by both or moved across another object breaks and must be replaced. Objects join permanently wherever they first meet. Partial tangrams can be broken and rebuilt from spare parts.

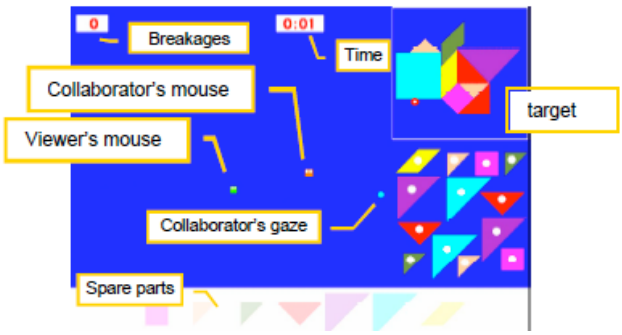

Figure 1. Joint Construction Task screen at trial onset.

Each player wears a head-mounted eye-tracker and a microphone and works on his/her own yoked screen. The system provides a full account of locations and movements of individual parts, constructed objects, and cursors for alignment with speech.

Participants, design and materials. Sixty-four Edinburgh University students worked in 32 same-sex pairs who had never met before. Three communication modalities were varied factorially within pairs: ability to speak while working, cross-projection of each player's current eye-track onto the other's screen, and cross-projection of mouse cursors. Non-speaking trials, for a different experiment, are not analyzed here. Players always saw their own mouse cursor and all parts. Players in 16 pairs were randomly assigned roles. The Manager was responsible for maximizing speed and accuracy, minimizing cost in breakages, and signaling the completion of each trial. The assistant was to help. Pairs with no roles otherwise had the same working instructions.

Table 1 Accessibility Coding Scheme (Ariel, 1990; Bard & Aylett, 2004)

Level	Definition	Examples
0	Indefinite NP	*a purple one* *one of the nearest blue pieces*
	Bare nominal	*pink one* *triangles*
1	Definite NP	*the red bit* *the other purple one*
2	Deictic NP	*those two little kids.*
	Deictic / Possess } Pron	*these* *mine*
3	Other Pronouns	*it*
	Clitic/inaudible.	*-/z/*

Coding. Coders used video and audio recordings to transcribe dialogues, time-stamp onset and offset of each referring expression and code each both for any on-screen object it referred to and for accessibility (Table 1). Codings were cross-checked by another coder and reviewed by one author. Based on explicit linguistic form, accessibility coding yields negligible disagreement between coders.

Results

Multinomial logistic regressions tested effects of cross-projection (of mouse or gaze cursors) and of local movements (moving the named part and 'hovering' - superimposing a mouse on a part without grasping or moving it) on distribution of 1775 first mentions over accessibility levels. Because the context classically calls for level 0, indefinite specific reference (*a red square*), effects reflect changes in the ratio of level 0 to other forms.

Results support the Declarative position. Speakers were not sensitive to what they could see of the collaborator's gaze or mouse actions. Instead, speaker actions affected form of referring expression. Helpfully, both role groups changed their introductory mentions while indicating the referent in ways the collaborator could see: moving the referent increased the rate of deictics (Mouse hidden (n = 939): B = -0.814, Wald statistic = 4.78, $p < .05$; Mouse visible (n = 836): B = -0.722, Wald = 5.05, $p < .05$). As the left panel of Figure 2 shows, hovering a visible, cross-projected mouse over the referent also increased the proportion of deictics globally (by decreasing definites B = 1.264 Wald = 6.95, $p < .01$). As the right panels of Figure 2 show, however, Manager-Assistant and No Role pairs differed (χ^2 = 276.00, Wald = 7.42, $p < .05$) when the collaborator could not see the mouse cursor. Only the former unhelpfully reflected private gestures, shifting from indefinites (22% v 8%) to deictics (25% v 40%) (Speaker hover x roles: B = 1.275, Wald = 4.99, $p < .05$) and definites (36% v 48%) (B = 1.075, Wald = 4.99, $p < .05$) when their mouse hovered over the named part. Thus for Manager-Assistant pairs alone, both public and private gestures brought changes in referring expressions.

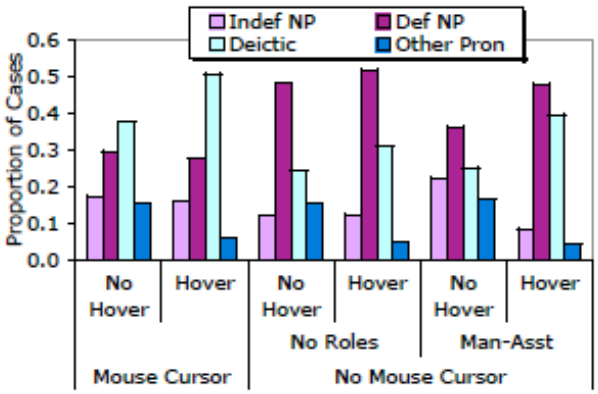

Figure 2. Accessibility of first mentions: effects of hovering mouse over referent, by cross-projection of mouse cursor to collaborator.

Alignment of attention

Though these results show a Declarative pattern, they may not be to the point. They come from dialogue situated in a rich visual context where speakers had to coordinate their attention in order to perform delicate joint physical actions. If referents are already in focus because they are visible (Smith, Noda, Andrews, & Jucker, 2005) or selected for joint attention by an ongoing task, then any form of referring expression could draw a listener's attention to the correct referent (Gundel, Hedberg, & Zacharski, 1993). With co-presence so restrictive, speakers could not fail to serve listeners' genuine needs.

To determine whether players' views of referents are so coordinated, we investigated the alignment of their gaze. People look at what they are talking about (Griffin, 2004; Meyer, van der Meulen, & Brooks, 2004), hear mentioned (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995), or anticipate will be mentioned (Altmann & Kamide, 2007). If task-situated dialogue coordinates attention, players should regularly look at the same things, regardless of syntactic elaboration or apparently egocentric design. But if referring expressions do have a problem to solve in joint action seetings, and are keyed to the genuine accessibility of referents, then we should find a difference between the No Role dyads, who referred cooperatively, and the Manager-Assistant pairs, who did not. The No Role dyads should show better alignment of attention and better keying of attention to referential form.

Method: Cross-recurrent gaze

Cross-recurrence analysis (Richardson & Dale, 2005; Richardson, Dale, & Kirkham, 2007) was applied to the eye-track records of the players whose introductory mentions are described above. This technique (Zbilut, Giuliani, & Webber, 1998) measures both absolutely and nearly simultaneous entrained activity. The regions of interest (ROI) for gaze were both fixed (the timer, target tangram, etc) and dynamic (movable parts and partly built tangrams). Fixations on blank areas of the background, looks off-screen and blinks were excluded. The measurement period was centered at the onset of the referring expression, using only those 936 expressions which were separated from the previous referring expression by at least 4s. Each player's gaze was located at increments of 20ms before being pooled into bins of 200ms. With one player's gaze location as a reference, the other player's gaze at each bin was checked for matches in ROI. The likelihood of spatial overlap between participants' eye movements was therefore examined when offset in time by up to 4s.

Table 2. Analyzed Referring Expressions by Level and Roles

	Coded Accessibility			
Roles Assigned	0	1	2	3
No Role	46	168	152	57
Manager-Assistantt	89	162	187	75

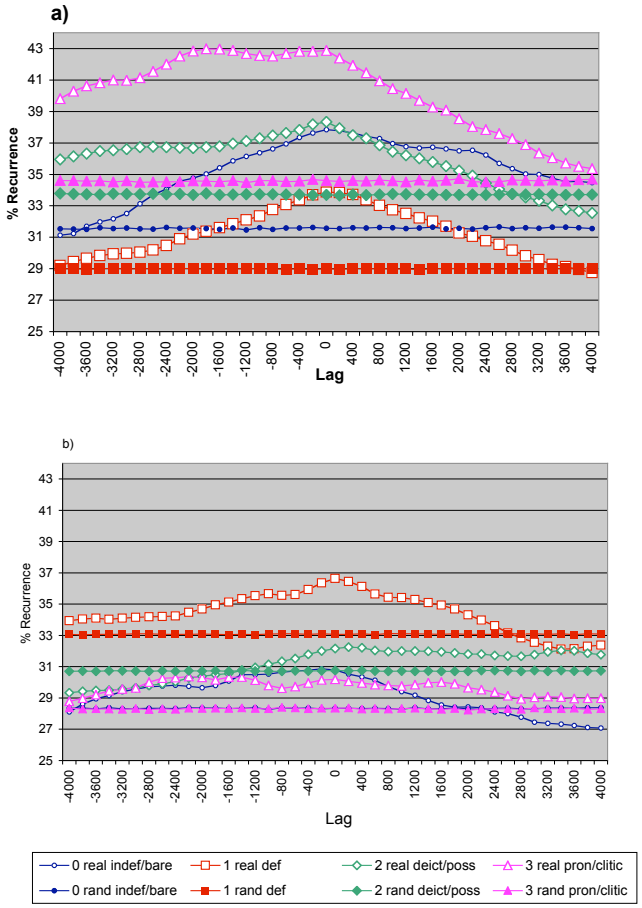

Figure 3. Real (hollow points) and randomized (filled points) cross recurrent gaze for a) No Role dyads b) Manager-Assistant dyads

Results

Table 2 gives the numbers of referring expressions at each accessibility level which could be analyzed for gaze alignment. Cross-recurrence results (Figures 3a and b) aggregate over all measurements in a sliding window. The y-axis shows percent recurrence, the proportion of all fixations within a bin which match between players. The x-axis shows the lag between one player's gaze and the other's. Colors indicate accessibility levels. Hollow points show real cross-recurrence results, reflecting the temporal organization of coordinated gaze, with 0 on the lag axis representing simultaneous gaze at identical ROIs, and other points representing gaze at identical ROIs offset in time. Filled points in matching colors provide a baseline: they represent agreement between one player's accurately ordered fixations and a random reordering of the other player's fixations. The figures show that joint physical action on visible objects does not guarantee that speaker and listener will coordinate attention so well that referent accessibility is always at ceiling.

In fact, gaze was not universally aligned. Though the peaks of real cross-recurrence curves (ranging from 30-43% coinciding gaze) are certainly above chance for an array with up to 15 regions of interest, alignment of gaze is not consistent. Manager-Assistant pairs (Figure 3b), who had been speaker-centric, aligned gaze significantly less than No Role pairs (Figure 3a), who had respected listeners' needs (real cross-recurrence, $F2(1, 904) = 4.66$, $p < .04$; absolute compared to randomized baselines, $F2(40, 36160) = 1.55$, $p < .02$). No Role cross-recurrence curves have a tent-like structure, indicating maximum alignment at minimal lag, while Manager-Assistant curves curves have a much flatter shape.

Moreover, there were robust differences in gaze coordination in exactly the directions predicted by referring expression usage and accessibility theory. Speakers who match referring expressions to listeners' needs should use less elaborate expressions for objects that are already accessible to listeners and more elaborate expressions for those to which listeners' attention needs to be directed. If speakers are not using this system in designing referring expressions, as in the Manager-Assistant pairs, no such prediction would be made.

To test these predictions, we examined the shape of the real cross-recurrence curves in Figure 3. Here, the negative lags indicate that listener looked at an object before the speaker did, while positive lags indicate that the speaker looked first. To see how far each case predominated for expressions at each accessibility level, we subtracted recurrent gaze percentages at each positive (speaker-first) lag from those at the corresponding (listener-first) negative lag. Figure 4 shows the averaged outcomes. A negative value here means that listeners' gaze predominantly followed speakers'.

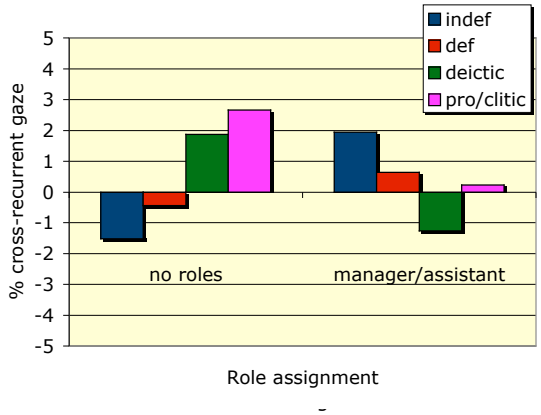

Figure 4. Order of players' gaze (listener first lags – speaker first lags) by referring expression type and role assignment.

As predicted, the relationship between referential form and gaze coordination was significant only for No Role players: the more accessible a form's referents should be, the stronger the tendency for the listener to look at ROIs

before the speaker (Linear mixed-effects regression: Coefficient = 2.91, p = .03; Spearman's *rho*: Coefficient =.10, p = .03 for No Role dyads; *n.s.* for Manager-Assistant dyads).

Discussion

Our analyses favor the Declarative views in several ways. First, the players who should have had more tendency to declare rather than negotiate were less affected by local circumstances. If form of referring expression were a reflex of either player's knowledge, the two role groups should behave alike in performing the same construction tasks. Second, elaboration of referring expressions was relevant to players' ability to coordinate attention in a demanding task. Under the same degrees of co-presence, those with more tendency to declare showed a weakness in aligning attention. Those with less emphasis on declaration showed good evidence of the underpinnings of accessibility theory, a relationship between the availability of a referent and the form of the expression used to describe it.

One question which remains is why, given the power to declare task strategy and accessibility, Manager-Assistant speakers choose to behave badly (Gerrig, Brennan, & Ohaeri, 2000). They are not being careless about their responsibilities. They seem to spend more time planning, for they spend more time looking off screen ($F(1, 501) = 14.31$, $p < .001$) or at the clock ($F(1, 501) = 6.39$, $p = .012$) (Doherty-Sneddon, Bruce, Bonner, Longbotham, & Doyle, 2002; Glenberg, Schroeder, & Robertson, 1998), though not enough more (2-3%) to make coordinated on-screen gaze impossible. They may simply be taking too individual a view of their own roles to consider the importance of the other player to their joint success. Or they may be relying on an official role relationship to assure that they can save effort in producing felicitous expressions by leaving their interlocutor to object if any expression is ultimately uninterpretable (Clark and Wilkes-Gibbs, 1986; Bard et al 2007).

What is clear is that even in highly situated dialogue, heavily contextualized and demanding assiduous visual attention, the relationship between reference and attention remains. Over several different forms of introductoy mentions, speakers choosing felicitous forms also adjust those forms as accessibility theory would predict, with more elaborate first mentions drawing in listeners' attention and less elaborate first mentions following after that attention has arrived.

Acknowledgments

This work was funded by EU Project JAST (FP6-003747-IP). The authors are grateful to the coders, to Tim Taylor, Craig Nicol, Joe Eddy, Jonathan Kilgour, and Jean Carletta for the experiment and analysis systems, and to J. P. de Ruiter for helpful discussions.

References

Altmann, G.T.M, & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language, 57*, 502-518.

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

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

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.

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

Bard, E. G., Anderson, A. H., 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. G., & Aylett, M. P. (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.

Brennan, S. E., Chen, X., Dickinson, C., Neider, M., & Zelinsky, G. (2007). Coordinating cognition: The costs and benefits of shared gaze during collaborative search. *Cognition*, 106, 1465-1477.

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

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

Brown-Schmidt, S., Campana, E., & Tanenhaus, M. K. (2004). Real-time reference resolution by naive participants during a task-based unscripted conversation. 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. 153–172). Cambridge, MA: MIT Press.

Carletta, J., Nicol, C., Taylor, T., Hill, R., de Ruiter, J. P., & Bard, E. G. (submitted). Eyetracking for two-person tasks with manipulation of a virtual world..

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

Doherty-Sneddon, G., Bruce, V., Bonner, L., Longbotham, S., & Doyle, C. (2002). Development of gaze aversion as disengagement from visual Information. *Developmental Psychology, 38*, 438-445.

Foster, M. E., Bard, E. G., 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 Human Robot Interaction, Amsterdam.

Gerrig, R., Brennan, S. E., & Ohaeri, J. O. (2000). What can we conclude from speakers behaving badly? . *Discourse Processes, 29*, 173-178.

Glenberg, A., Schroeder, J., & Robertson, D. (1998). Averting the gaze disengages the environment and facilitates remembering. *Memory & Cognition, 26*(4), 651-658.

Griffin, Z. M. (2004). Why look? Reasons for eye movements related to language production. In J. Henderson & F. Ferreira, (Eds.), The *integration of language, vision, and action: Eye movements and the visual world* (pp. 213-247). New York: Taylor and Francis.

Gundel, J. K., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language, 69*, 274-230.

Gundel, J. K., N, Hedberg and R, Zacharski. (2001). Definite descriptions and cognitive status in English: Why accommodation is unnecessary. *Journal of English Language and Linguistics, 5*, 273-295.

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

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

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

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition, 89*, 25–41.

Kranstedt, A., Lücking, A., Pfeiffer, T., Rieser, H., & Wachsmuth, I. (2006). Deictic Object Reference in Task-oriented Dialogue. In G. Rickheit & I. Wachsmuth (Eds.), *Situated Communication,* (pp. 155-207). Berlin: Mouton de Gruyter.

Louwerse, M., Dale, R., Jeuniaux, P., Benesh, N., Watanabe, S. & Bard, E. G. (2009). Mimicry in action: When people imitate in face-to-face conversation. Paper presented at the annual meeting of the Society for Text and Discourse.

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

Richardson, D. C., & 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. C., 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.

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

Steedman, M. (2007). Information-structural semantics for English intonation. In C. Lee, M. Gordon & D. Büring (Eds.), *Topic and Focus: Cross-linguistic Perspectives on Meaning and Intonation,* Vol. 82, (pp. 245-264). Dordrecht: Kluwer.

Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M., & Sedivy, J.C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*, 1632-1633.

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